Analytisches CRM: Anwendungen (Ü)

Sommersemester 2014

Dipl.-Math. Josef Bauer
Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Lehrstuhl für ABWL und Wirtschaftsinformatik

Katholische Universität Eichstätt-Ingolstadt

Organisatorisches

- **Klausur**
  - **Ersttermin (08.07.2014): Klausur 30 min (2/3) + Case (1/3) (innovativ)**
  - **Zweittermin (30.09.2014): Klausur 30 min (2/3) + Case (1/3) (innovativ) / 45 min (regulär)**
  - **Case: Abgabe am 15.07.2014 12 Uhr** auf CD in den Briefkasten des Lehrstuhls
  - **Klausurrelevant:** Alle behandelten Inhalte des Skripts, Workshop Altran, ggf. Exkurse
- **Zeitplan**

Tag	Kapitel	Name
1	1-4	IBM SPSS Modeler: Grundlagen, Knoten-Übersicht, Einlesen von Daten
2	5	Datenaufbereitung
3	6	Data Mining-Modelle: Künstliche NN, Entscheidungsbäume, Regression
	6	Clustering, Assoziationsanalyse
4	6	Workshop ALTRAN CIS (Herr Tobias von Martens)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Kleine Tipps zum Vorankommen

- **Streams und Daten nach jeder wichtigeren/umfangreicheren Veränderung sichern!**
- **Keine Angst vor Trial and Error!**
  - Beim selbstständigen Ausprobieren, Versuchen und Untersuchen von Einstellungen, die von der Übung abweichen, lernt man am meisten.
  - Dazu am besten die Übungsdaten kopieren und für T&E mit den kopierten Dateien arbeiten (dann kann auch nichts kaputt gehen)
- **Bei Fragen immer fragen!** ☺
- **Pfad für die Übungsdaten:**
  M:\Win\WI_Infos\aCRM_Übung_SS14\Daten_Übung

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

1. Analytisches CRM
2. IBM SPSS Modeler
3. Knoten-Übersicht
4. Einlesen von Daten
5. Datenaufbereitung
6. Data Mining-Modelle

OLAP und Data Mining: Unterschiedliche Fragestellungen

Fragestellung	Data Mining	OLAP
Kundenwert	Welche 10 Kunden bieten uns das größte Deckungsbeitrags*potenzial*?	Wer waren *letztes Jahr* unsere 10 besten Kunden?
Kündigeranalysen	Welche Kunden werden *wahrscheinlich* in den nächsten 6 Monaten zu einem Konkurrenten wechseln?	Welche Kunden haben *letzten Monat* gekündigt?
Cross-Selling	Welche unserer bestehenden Kunden werden am *wahrscheinlichsten* unser neues Produkt kaufen?	Wie viele Einheiten unseres neuen Produktes *haben* wir an unsere bestehenden Kunden *verkauft*?
Kundenakquise	Wie ist das Profil der Leute, die auf unsere zukünftigen Mailings am *wahrscheinlichsten* antworten werden?	Wie *war* die Responsequote unseres Mailings?
Kreditrisiko-Management	Hat dieser Kunde *wahrscheinlich* eine gute Bonität? Wird sich die Bonität ändern?	Welche Kunden *waren* zahlungsunfähig?

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Data Mining-Analysen

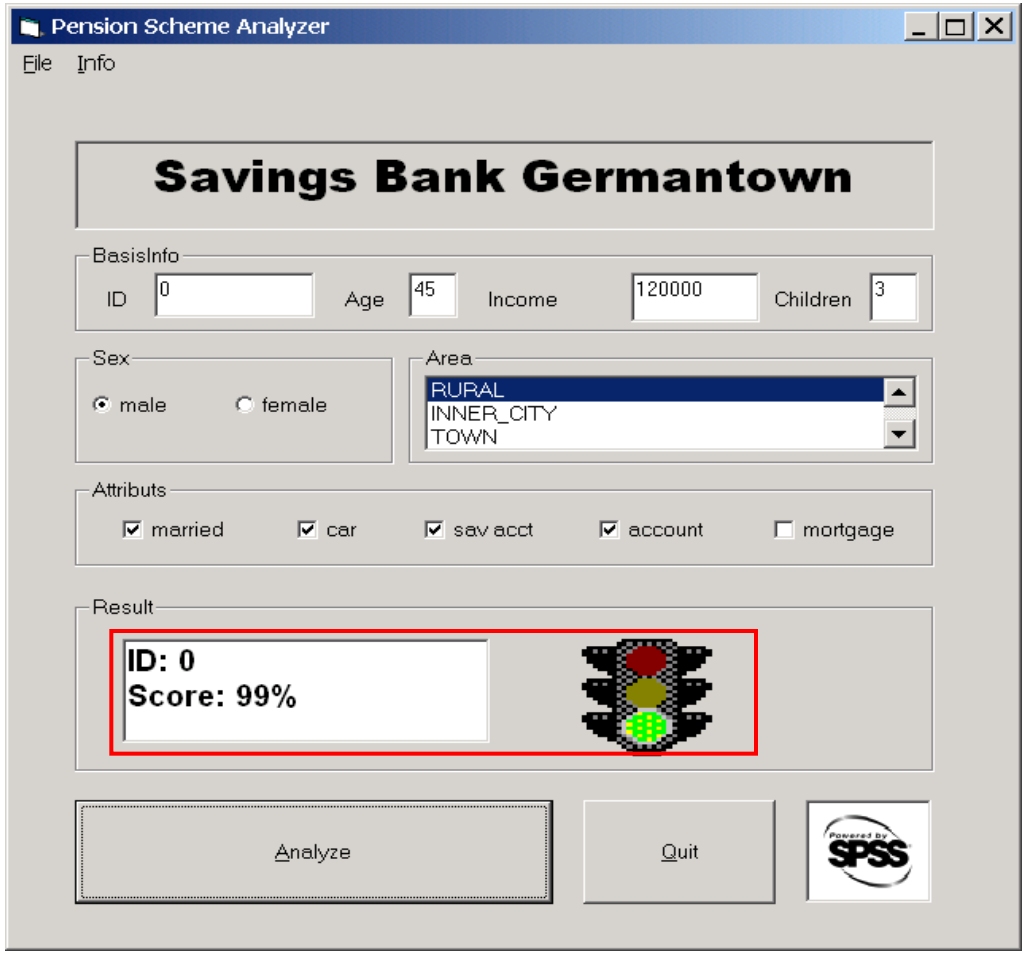
- Automatisches Durchsuchen großer Datenmengen und Erkennen von nützlichen Mustern.
- **Beispiele:**
  - Erstellen von Kundenprofilen
  - Ermitteln von Kundenkapitalwerten
  - Erkennen und Vorhersagen von Betrugsfällen
  - Ermitteln und Vorhersagen von Sequenzen in Website-Daten
  - Vorhersagen von zukünftigen Trends bei Verkaufszahlen
  - Zielgruppenselektion für Mailingaktionen
  - Einschätzung von Bonität bzw. Kreditrisiko
  - Vorhersagen der Kundenverlustrate
  - Klassifizieren von Kunden

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Praxisbeispiel für Data Mining-Anwendung

Data Mining-Modell für die Gewährung eines Altersvorsorgeplans einer Bank

- Ein Modell wird aufgrund historischer Merkmalsausprägungen in Kundendaten trainiert (= entwickelt)
- Über sog. Solution Publishers kann das vorab erstellte Modell allen Mitarbeitern (z. B. im CIC) in der Anfrageabwicklung operativ zur Verfügung gestellt werden.



Weitere Beispiele für Scorings

- Risikoanalyse bei Kreditvergabe in Banken
- Vorhersagesystem für Airline, um Überbuchungen zu vermeiden

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

1. Analytisches CRM
2. IBM SPSS Modeler
3. Knoten-Übersicht
4. Einlesen von Daten
5. Datenaufbereitung
6. Data Mining-Modelle

Übersicht: Data Mining-Software

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [1880 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools	% users in 2013 % users in 2012
Rapid-I RapidMiner/RapidAnalytics free edition (737), 30.9% alone	39.2% 26.7%
R (704), 6.5% alone	37.4% 30.7%
Excel (527), 0.9% alone	28.0% 29.8%
Weka / Pentaho (269), 5.6% alone	14.3% 14.8%
Python with any of numpy/scipy/pandas/iPython... packages (250), 0% alone	13.3% 14.9%
Rapid-I RapidAnalytics/RapidMiner Commercial Edition (225), 52.4% alone	12.0%
SAS (202), 2.0% alone	10.7% 12.7%
MATLAB (186), 1.6% alone	9.9% 10.0%
StatSoft Statistica (170), 45.9% alone	9.0% 14.0%
IBM SPSS Statistics (164), 1.8% alone	8.7% 7.8%
Microsoft SQL Server (131), 1.5% alone	7.0% 5.0%
Tableau (118), 0% alone	6.3% 4.4%
IBM SPSS Modeler (114), 6.1% alone	6.1% 6.8%
KNIME free edition (110), 1.8% alone	5.9% 21.8%
SAS Enterprise Miner (110), 0% alone	5.9% 5.8%
Rattle (84), 0% alone	4.5%
JMP (77), 7.8% alone	4.1% 4.0%
Orange (67), 13.4% alone	3.6% 5.3%
Other free analytics/data mining software (64), 3.1% alone	3.4% 4.9%
Gnu Octave (54), 0% alone	2.9%
Revolution Analytics R Enterprise (53), 1.9% alone	2.8% 1.4%
Predixion Software (51), 43.1% alone	2.7% 0.4%

Enterprise Miner

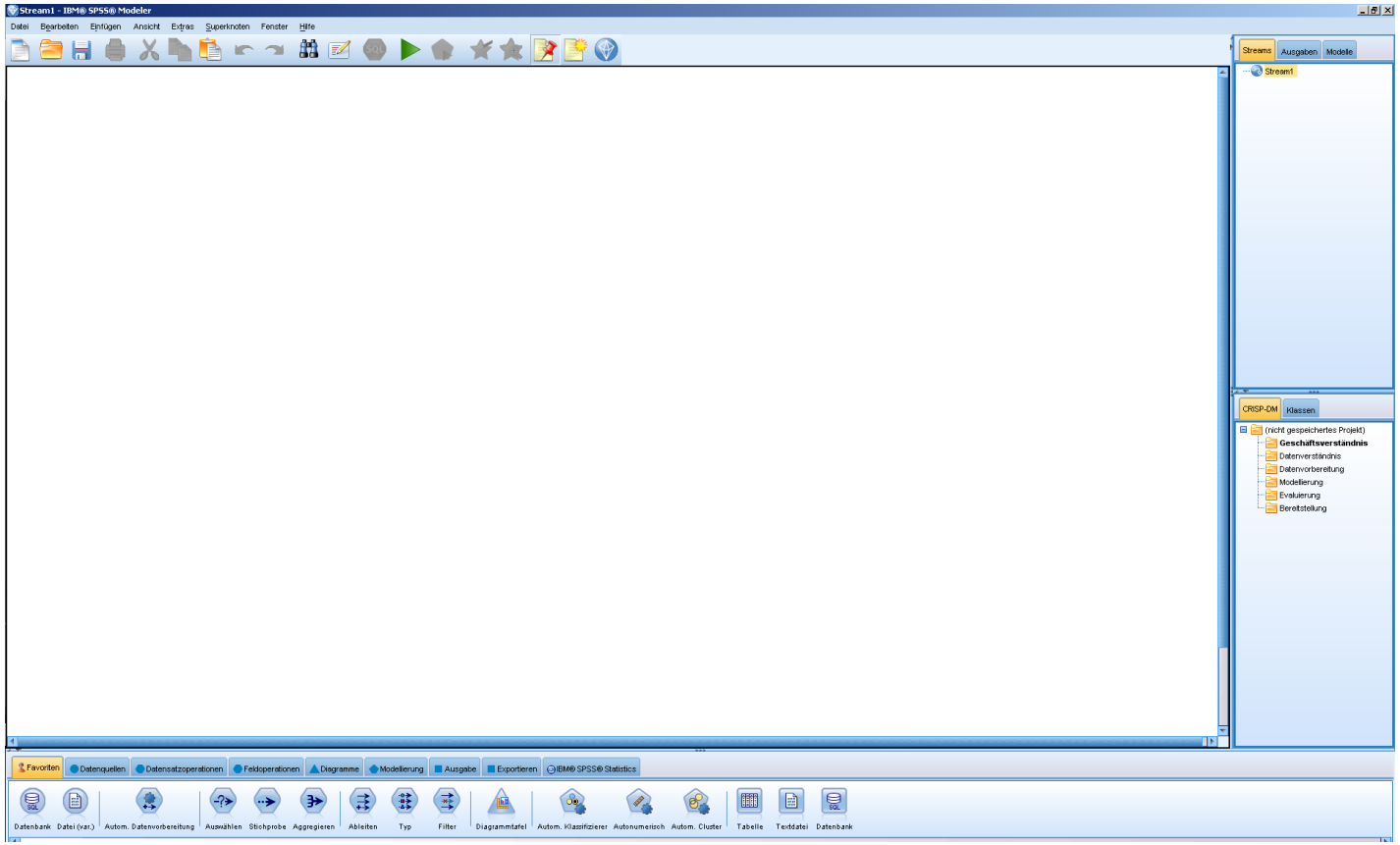
Quelle: http://www.kdnuggets.com/polls/2013/analytics-big-data-mining-data-science-software.html (Zugriff: 11.06.2013)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Modeler starten

- Wählen Sie im **Windows-Startmenü**:

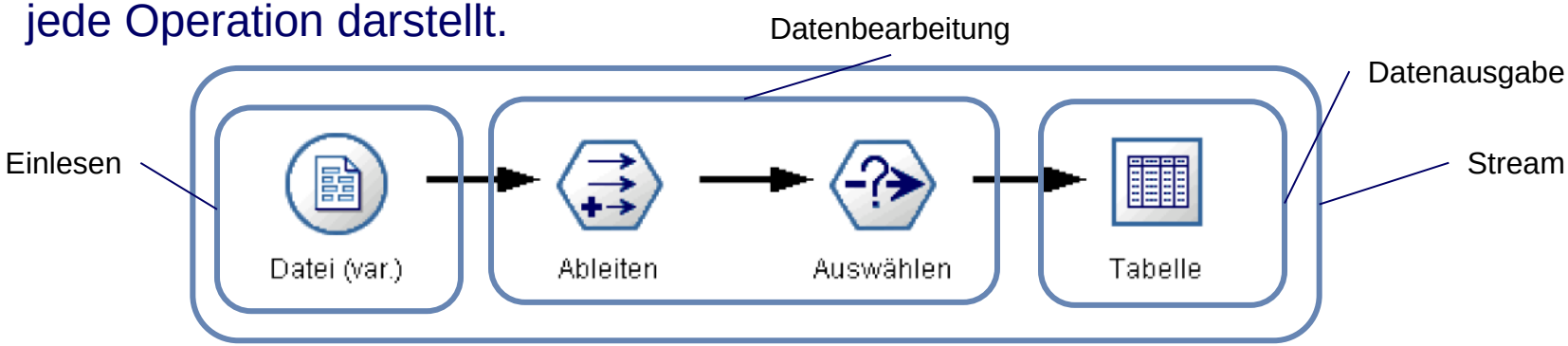
  (Alle) Programme
  IBM SPSS Modeler
  IBM SPSS Modeler


Stream1 - IBM® SPSS® Modeler
Datei Bearbeiten Einfügen Ansicht Extras Superknoten Fenster Hilfe
Streams Ausgaben Modelle
Stream1
CRISP-DM Klassen
(nicht gespeichertes Projekt)
Geschäftsverständnis
Datenverständnis
Datenvorbereitung
Modellierung
Evaluierung
Bereitstellung
Favoriten Datenquellen Datensatzoperationen Feldoperationen Diagramme Modellierung Ausgabe Exportieren IBM® SPSS® Statistics
Datenbank Datei (var.) Autom. Datenvorbereitung Auswählen Stichprobe Aggregieren Ableiten Typ Filter Diagrammtafel Autom. Klassifizierer Autonumerisch Autom. Cluster Tabelle Textdatei Datenbank

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Modeler auf einen Blick

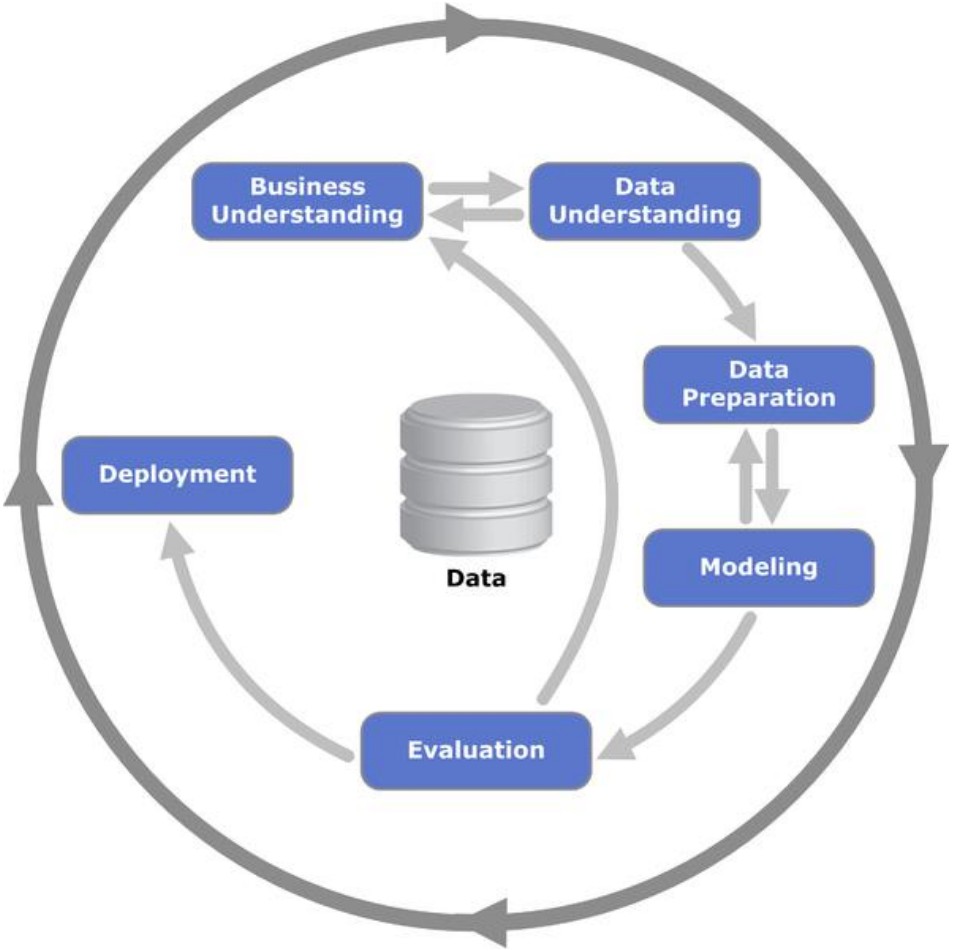
- Arbeiten im Modeler bedeutet **Arbeiten mit Daten**.
  In der einfachsten Form besteht die Arbeit im Modeler aus drei Schritten nach dem EVA-Prinzip:
  1. Einlesen der Daten in den Modeler
  2. Bearbeiten der Daten
  3. Daten an ein Ziel senden
- Diese Reihenfolge wird als **Daten-Stream** bezeichnet, da die Daten Datensatz für Datensatz von der Quelle durch jeden Bearbeitungsschritt zum Ziel fließen.
- Jede Modeler-Operation wird durch einen **Knoten** (Symbol) dargestellt.
- Miteinander verbundene Knoten bilden einen **Stream**, der den Datenfluss durch jede Operation darstellt.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Das CRISP-DM-Prozessmodell im Modeler

- CRISP-DM 1.0



Business Understanding
This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

Data Understanding
The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation
The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling
In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

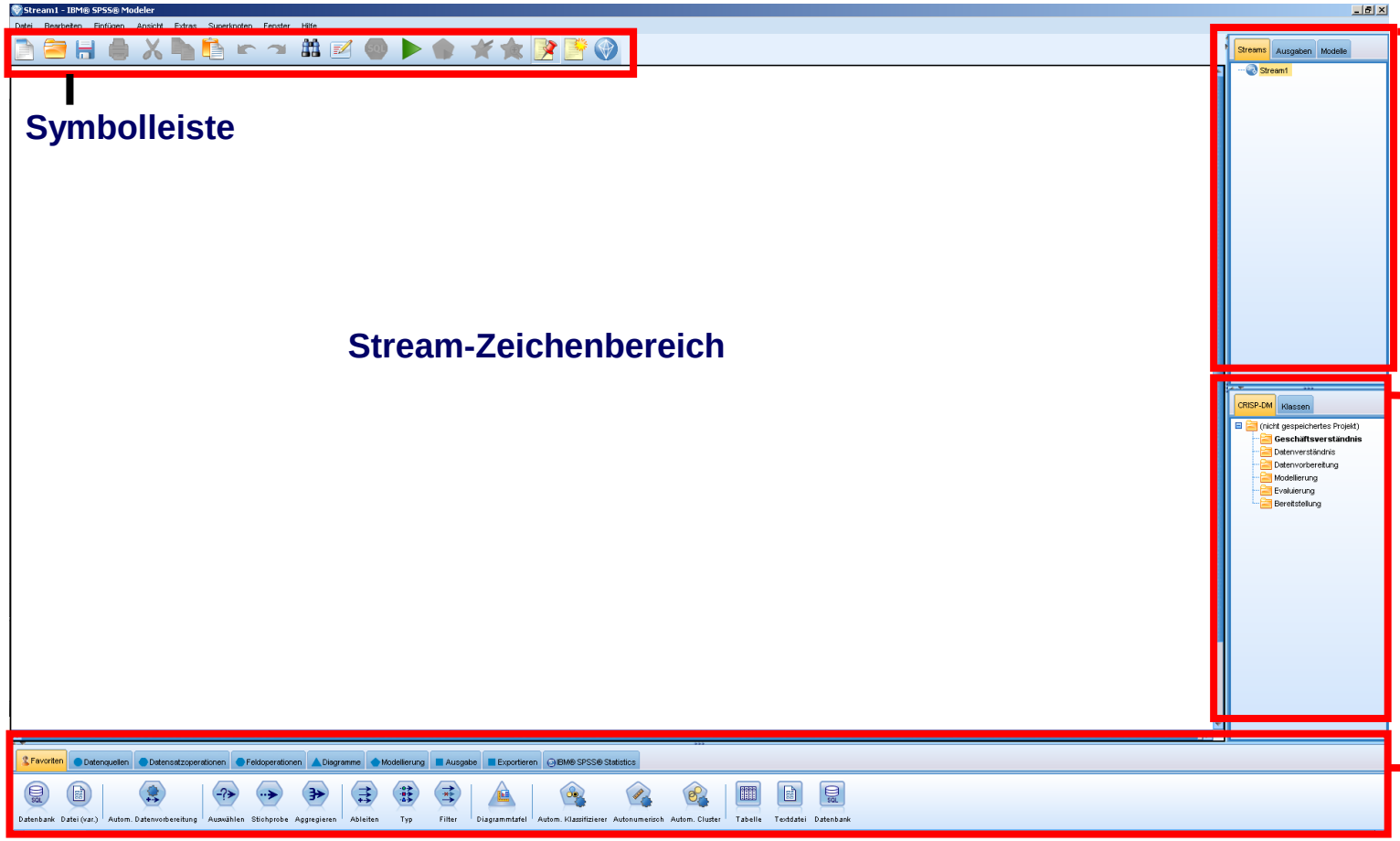
Evaluation
At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment
Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models.

Quelle: http://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png (Zugriff: 11.06.2013)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Die Arbeitsoberfläche



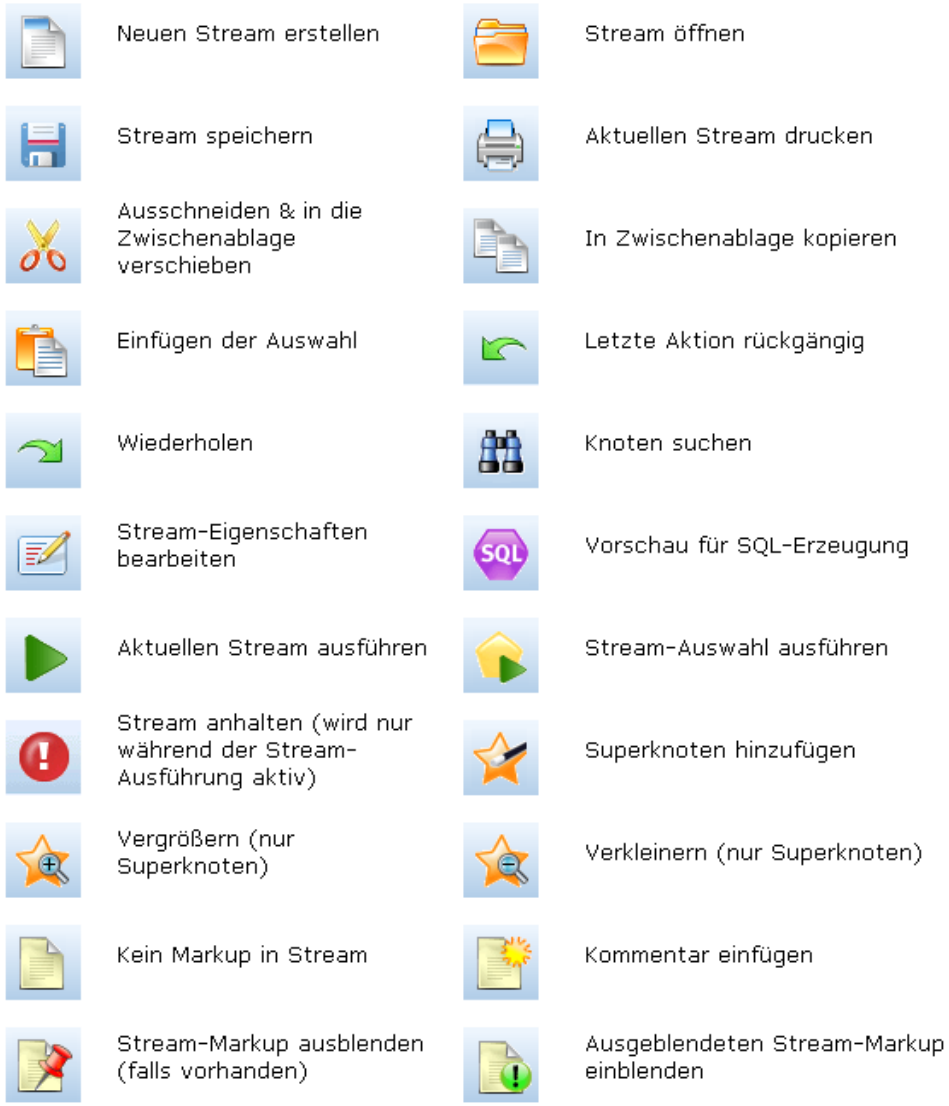
Manager zum Anzeigen / Verwalten von z. B. Streams, Diagrammen, Modellen

CRISP-Projekt-Fenster mit Streams, Ausgaben, Dateien, etc. zum gegenwärtigen Data Mining Projekt

Paletten mit jeweils einer Gruppe verwandter Knoten

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Die Symbole der Symbolleiste



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Verwenden der Maus

- **Einfaches Klicken der linken/rechten Maustaste:**
  - Optionen aus Menüs auswählen
  - Kontextsensitive Menüs öffnen
  - Verschiedene andere Standardsteuerelemente und Optionen verwenden
- **Drücken mit Halten der linken Maustaste:**
  - Knoten verschieben bzw. ziehen (im Zeichenbereich, aus Palette in Zeichenbereich)
- **Doppelklicken der linken Maustaste:**
  - Knoten auf dem Stream-Zeichenbereich ablegen (aus Palette)
  - Optionen bereits vorhandener Knoten bearbeiten
- **Mittlere Maustaste (bzw. Scrollrad):**
  - Gedrückt halten und Cursor ziehen, um Knoten im Stream-Zeichenbereich miteinander zu verbinden

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

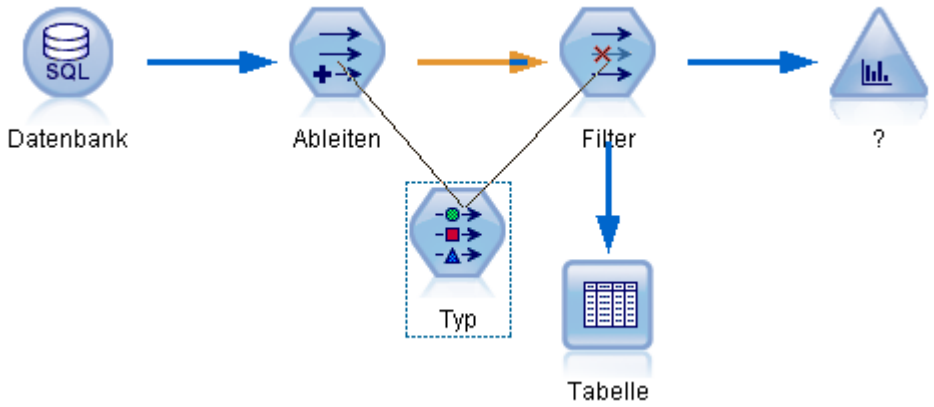
Arbeiten mit Knoten 1/3

- **Knoten auf dem Stream-Zeichenbereich ablegen**
  - Doppelklick auf das Knoten-Symbol. Knoten wird automatisch mit dem aktuellen Stream verbunden.
  - Knoten per Drag-and-Drop aus der Palette in den Stream-Zeichenbereich ziehen.
  - Markieren des gewünschten Knotens im Paletten-Fenster und ablegen des Knotens durch einfachen Mausklick an der entsprechenden Stelle im Stream-Zeichenbereich.
- **Knoten verschieben**
  - Entsprechenden Knoten im Stream-Zeichenbereich markieren und bei gedrückter linker Maustaste verschieben.
- **Knoten verbinden**
  - Doppelklick auf das Knoten-Symbol in der entsprechenden Palette (s. o.)
  - Ausgangsknoten markieren und bei gedrückter mittlerer Maustaste mit nachfolgendem Knoten verbinden.
  - Ausgangsknoten markieren, im Kontextmenü (rechte Maustaste) „Verbinden…“ wählen und mit nachfolgendem Knoten verbinden.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Arbeiten mit Knoten 2/3

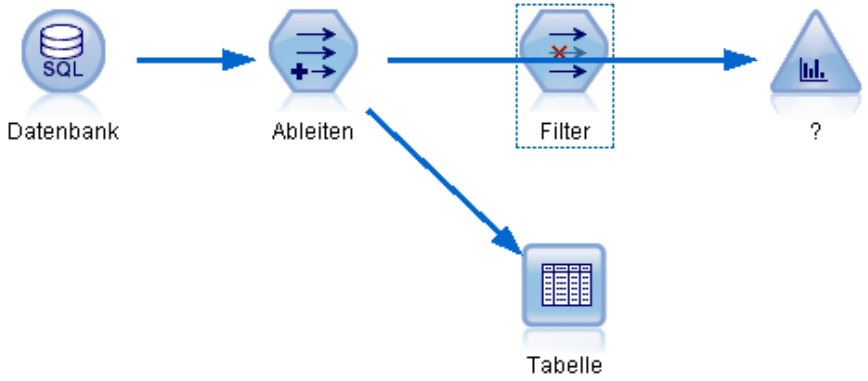
- **Knoten löschen**
  - Knoten markieren und die Taste „Entf“ drücken.
- **Knoten-Verbindung löschen**
  - Mit der rechten Maustaste auf den Pfeil klicken und „Verbindung löschen“ wählen.
  - Knoten markieren und F3 drücken.
- **Hinzufügen von Knoten in bereits vorhandene Verbindungen**
  - Mit der linken oder mittleren Maustaste den ursprünglichen Verbindungspfeil auf den neuen Knoten ziehen.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Arbeiten mit Knoten 3/3

- **Knoten umgehen**
  - Um einen zuvor eingebundenen Knoten in einem Stream auszulassen, den entsprechenden Knoten markieren und Doppelklick auf die mittlere Maustaste (alternativ: ALT-Taste gedrückt halten und Doppelklick auf die linke Maustaste).



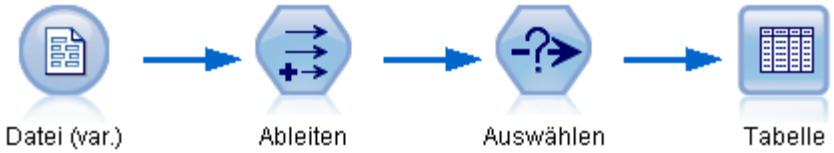
- **Knoten bearbeiten** (Umbenennen, Optionen auswählen, Löschen etc.)
  - Mit der rechten Maustaste auf den Knoten klicken und Menüpunkt wählen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 2.1

Einarbeitung in die Modeler-Arbeitsumgebung

Erstellen Sie den abgebildeten Stream.
(Hinweis: Knoten- bzw. Stream-Optionen müssen nicht verändert werden)

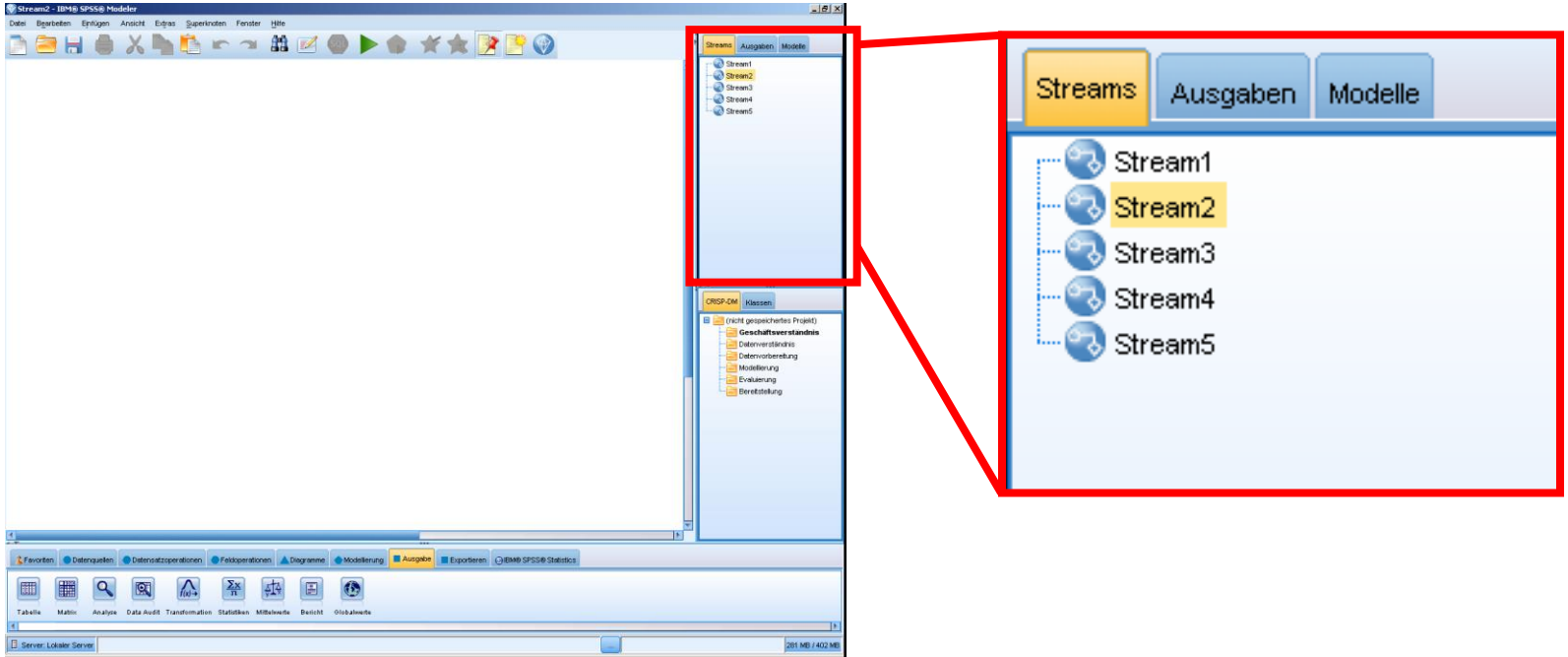


Der abgebildete Stream wird bspw. verwendet, um:
- eine Datenquelle zu öffnen
- ein neues Datenfeld hinzuzufügen
- Datensätze basierend auf den Werten im neuen Feld auszuwählen
- das Ergebnis in einer Tabelle anzuzeigen

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Arbeiten mit Streams

- Wenn Quell-, Prozess- und Endknoten im Stream-Zeichenbereich verbunden sind wurde ein **Stream** erstellt. Als Knotensammlung können Streams gespeichert, mit Anmerkungen versehen und zu Projekten hinzugefügt werden.
- Im Modeler können **mehrere Daten-Streams geöffnet, bearbeitet und verwendet** werden. Die rechte Seite des Modeler-Fensters enthält das Manager-Tool, mit dem man durch die aktuell geöffneten Streams navigieren kann.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Ausführen von Streams

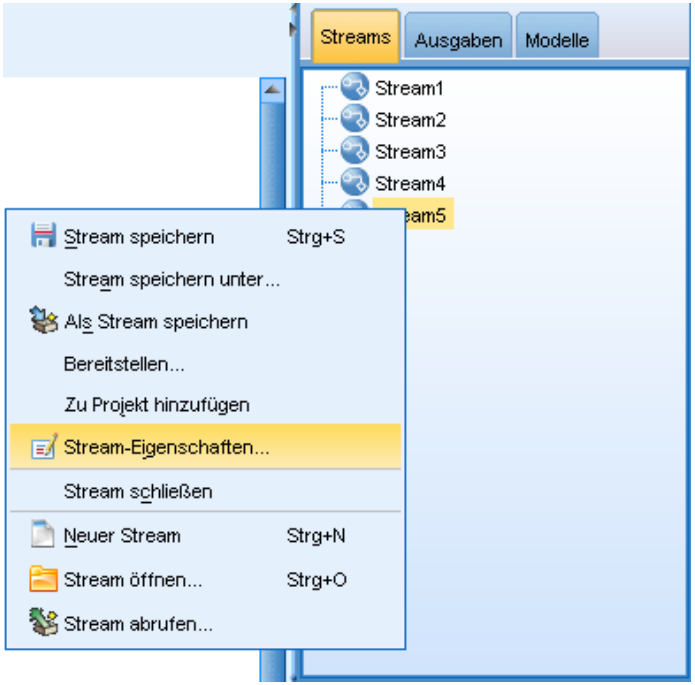
- Einen Stream auszuführen bedeutet, die **Daten „durch die Knoten im Stream fließen zu lassen“.** Dazu gibt es verschiedene Möglichkeiten:
- **Gesamten Daten-Stream ausführen:**
  - Auf die Ausführungsschaltfläche in der Symbolleiste klicken.
  - Im Menü „Extras“ die Option „Ausführen“ auswählen.
- **Teil eines Daten-Streams ausführen:**
  - Endknoten im Stream-Zeichenbereich markieren und auf die Schaltfläche in der Symbolleiste klicken.
  - Mit rechter Maustaste auf Endknoten im Stream-Zeichenbereich klicken und im Kontextmenü „Ausführen“ auswählen.
  - Mit linker Maustaste auf Endknoten im Stream-Zeichenbereich klicken und Tastenkombination „Strg + e“ drücken.
- **Ausführung eines gerade laufenden Streams anhalten:**

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Stream-Eigenschaften

Zum Einstellen der Stream-Eigenschaften

- Symbol aus der Symbolleiste aufrufen
- den Menüpunkt *Extras – Stream-Eigenschaften – Optionen...* aufrufen
- oder mit der rechten Maustaste im Managerfenster auf den Stream klicken und im Kontext-Menü den Punkt „Stream-Eigenschaften…“ auswählen

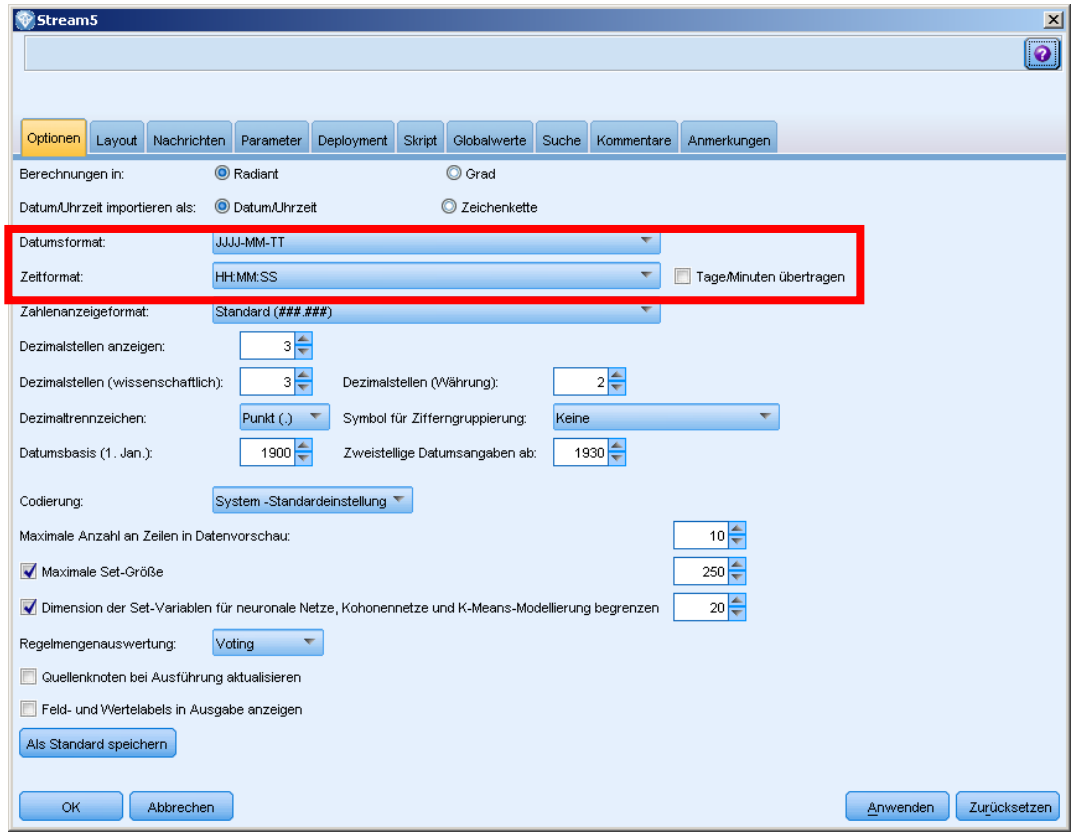


Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Stream-Eigenschaften

Datums- und Uhrzeit-Format

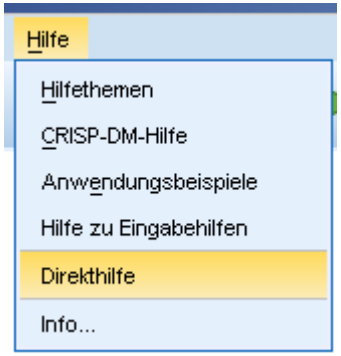
- Im Data Mining und insbesondere im Web Mining wird oft mit Feldern gearbeitet, die Datum und Uhrzeit enthalten.
- Der Modeler kennt viele Datums- und Uhrzeit-Formate, der Nutzer muss jedoch die entsprechenden Einstellungen auswählen.



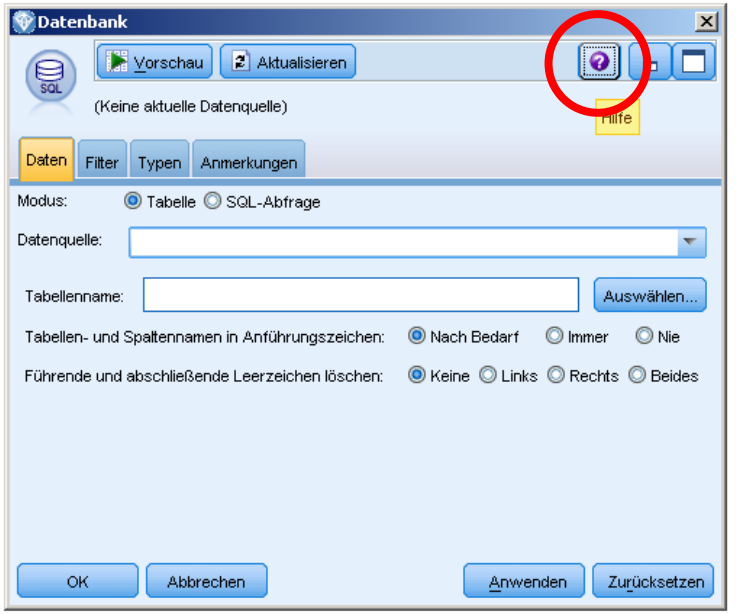
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Hilfe-Funktion

- **Direkthilfe**
  Um allgemeine Hilfe zu Knoten und Symbolleistenelementen aufzurufen, „Direkthilfe“ im Hilfemenü von Modeler auswählen. Der Cursor verwandelt sich in ein Fragezeichen, mit dem man auf ein beliebiges Element im Stream-Zeichenbereich oder in den Paletten klicken kann. Es öffnet sich ein Hilfefenster mit Informationen zum ausgewählten Element.



- **Kontextsensitive Hilfe**
  Auf Hilfeschaltfläche / Hilfesymbol in einem Dialogfeld klicken, um ein Hilfethema speziell für die Steuerelemente in diesem Dialogfeld aufzurufen.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Dokumentation

- Sehr umfangreiche Modeler-Dokumentation erreichbar über den Menüpunkt „Hilfe“.
  - **Modeler-Benutzerhandbuch**
  - **Knoten-Referenz**
  - Skripterstellungs-, Automatisierungs- und CLEM-Referenz
  - In-Database Mining-Handbuch
  - Server- und Leistungshandbuch
  - Anwendungshandbuch
  - Algorithmushandbuch
  - Modeler Solution Publisher
  - **CRISP-DM 1.0 Handbuch**
  - SPSS-Befehlssyntaxreferenz
  - Online-Hilfe und Lernprogramme

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 2.2

Zurechtfinden in der Modeler-Hilfe

- Finden Sie heraus, welche Optionsparameter sich beim Knoten "Datei (var.)" einstellen lassen.
- Wozu dient der Knoten "Data Audit"?

- Was ist CLEM?
- Suchen Sie in den Dokumentationsunterlagen nach einer Abbildung des CRISP-DM-Prozessmodells.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

1. Analytisches CRM
2. IBM SPSS Modeler
3. Knoten-Übersicht
4. Einlesen von Daten
5. Datenaufbereitung
6. Data Mining-Modelle

Datenquellen

Analytic Server: Ermöglicht die Ausführung eines Streams in HDFS (Hadoop Distributed File System)

Datenbank: Datenimport aus Datenbanken über ODBC

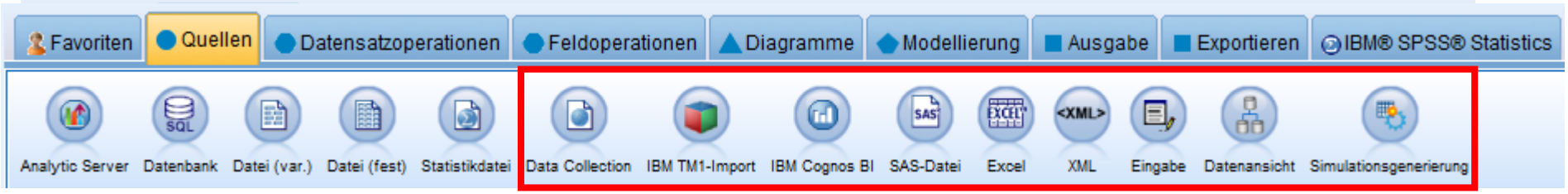
Datei (var.): Datenimport aus Textdateien mit Feldern variabler Größe

Datei (fest): Datenimport aus Textdateien mit Feldern fester Größe

Statistics-Datei: Importieren von SPSS-Dateien

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenquellen



Data Collection: Importieren von Umfragedaten auf der Grundlage des Data Collection Model von SPSS

IBM TM1-Import: Importieren von Daten aus IBM TM1

IBM Cognos BI: Importieren von Daten aus IBM Cognos BI Datenbanken

SAS-Datei: Importieren von SAS-Dateien

Excel: Importieren von Excel-Dateien

XML: Importieren von XML-Dateien

Eingabe: Manuelle Eingabe von Feldern und Werten

Datenansicht: Lesen von Daten aus einer Repository-Quelle

Simulationsgenerierung: Erzeugung simulierter Daten

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datensatzoperationen

Auswählen: Wählt Datensätze aus, die eine Bedingung erfüllen bzw. nicht erfüllen (Bsp.: Einkommen = "hoch")

Stichprobe: Stichprobenziehung, bspw. zur Reduzierung der Datenmenge

Sortieren: Datensätze anhand der Werte eines oder mehrerer Felder in aufsteigender oder absteigender Reihenfolge sortieren

Balancieren: Unausgeglichene Gruppenverteilung ausbalancieren (Over- bzw. Undersampling)

Duplikat: Doppelte Datensätze erkennen bzw. herausfiltern

Aggregieren: Aggregieren von Variablen nach Schlüsselvariablen (Bsp.: Durchschnittlicher Umsatz pro Segment)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datensatzoperationen

RFM-Aggregat: Aggregation von Datensätzen derselben ID über Aktualität, Häufigkeit und monetären Wert

Zusammenführen: Zusammenfügen verschiedener Dateien zu einer Datentabelle „nebeneinander“

Anhängen: Zusammenfügen verschiedener Dateien zu einer Datentabelle „untereinander“ (sinnvoll bei gleicher Feldstruktur der Dateien)

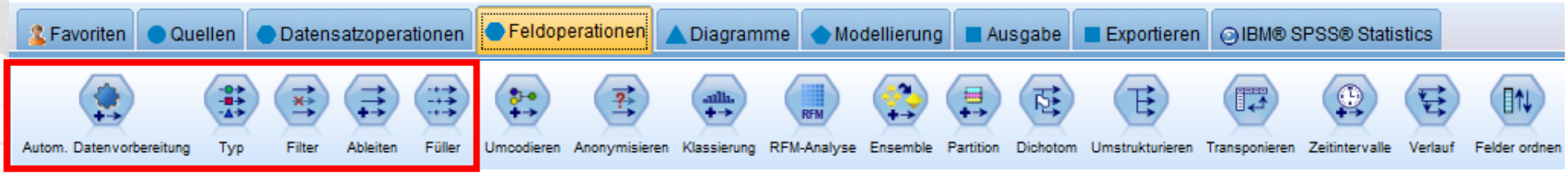
Streaming-ZR: Für Zeitreihenmodelle

R-Transformation: Datentransformation mit R

Space-Time-Boxes: Für Raum- und Zeitinformationen

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Feldoperationen



Automatische Datenvorbereitung:
Übernimmt selbstständig die Schritte der Datenvorbereitung

Typ: „Verwalten“ der Felder
(Skalenniveau, Wertebereich, Missing Values, Einstellung Eingabe-/Ausgabefeld)

Filter: Herausfiltern nicht mehr benötigter Datenfelder, Umbenennen von Feldnamen

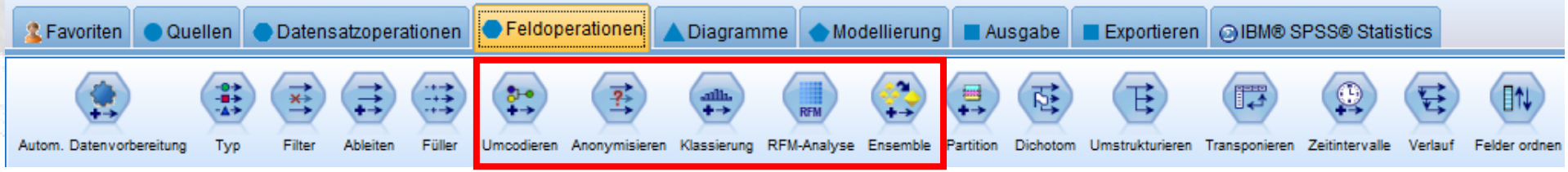
Ableiten: Neue Datenfelder aus bestehenden Daten ableiten

- Formel: „Normale“ Berechnung, z. B. Local + Longdist + ... oder über CLEM
- Flag: Dichotomisieren von Feldern auf Basis einer Bedingung
- Set: Kategorisieren von Feldern
- Status: Ein- und Ausschalten eines Wertes in Abhängigkeit einer Bedingung
- Anzahl: Erhöht den Wert des neuen Feldes, wenn die Bedingung erfüllt ist
- Bedingt: Bedingte Berechnung

Füller: Feldwerte ersetzen, um bspw. Nullwerte durch einen bestimmten Wert zu ersetzen

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Feldoperationen



Umcodieren: Transformation eines Sets diskreter Werte in ein anderes (Bsp.: Umkodieren der Werte des Datenfelds „Produkt“ in die drei Gruppen Möbel, Küchenzubehör, sowie Bad und Bettwäsche)

Anonymisieren: Verschleiern von Feldnamen und/oder -werten

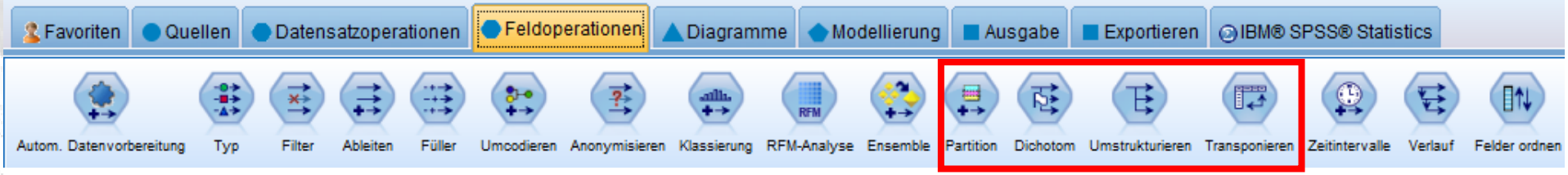
Klassierung: Kategorisieren von Variablen (z. B. Altersklassen, Einkommensklassen etc.)

RFM-Analyse: "RFM-Analyse" (Recency-, Frequency-, Monetary-Analyse)

Ensemble: Verbindung mehrerer Modellscores zu einem zusammengesetzten Score (sog. Meta-Modelling)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Feldoperationen



Partition: Zufälliges Aufteilen der Daten in Partitionen (z. B. für Training, Test und Validierung)

Dichotom: Generiert neue dichotome Felder (Flags) auf der Grundlage von Kategoriewerten
(z. B. fünf Einkommensklassen → neues Feld „HohesEinkommen“ mit den Werten 0/1)

Umstrukturieren: Generiert neue Felder auf der Grundlage von Kategoriewerten. Neue Felder müssen aber nicht die Werte 0/1 besitzen (wie bei Dichotomknoten), sondern können bspw. Kopien von Werten aus der Datentabelle erhalten.
(s. **Beispiel auf nächster Folie**)

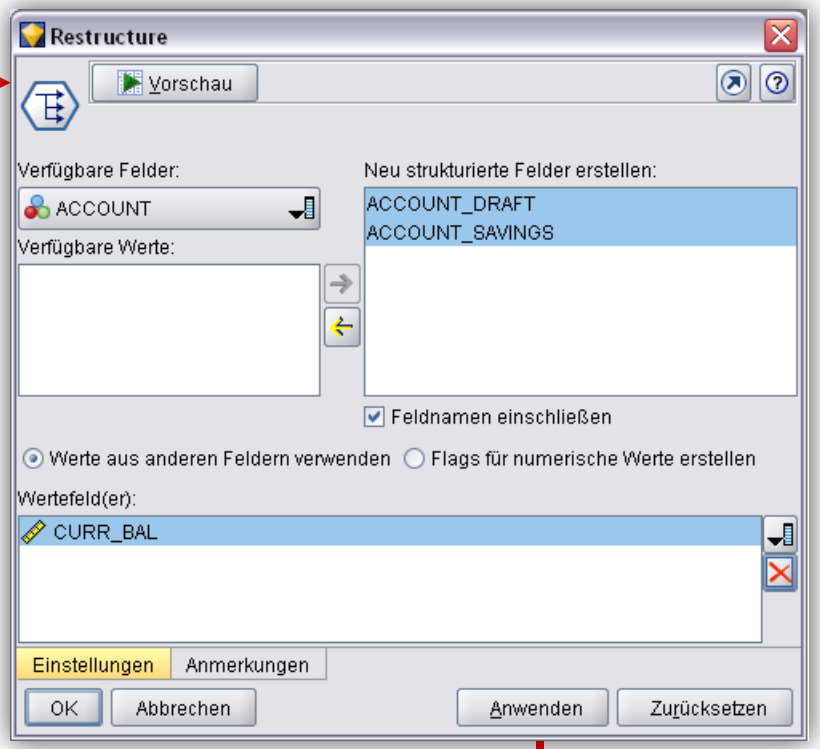
Transponieren: Vertauscht die Daten in Zeilen und Spalten

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Beispiel: Umstrukturierungs- & Aggregat-Knoten

Beispieldaten vor der Neustrukturierung

CustID	Account	Open_Bal	Current_Bal
12701	Draft	1000	1005.32
12702	Savings	100	144.51
12703	Savings	300	321.20
12703	Savings	150	204.51
12703	Draft	1200	586.32



Das ursprüngliche Daten-Set enthält ein Feld „Account“ mit den Werten Draft und Savings. Für jedes Konto werden der Anfangssaldo und der aktuelle Saldo festgehalten; einige Kunden besitzen mehrere Konten von jedem Typ. Mit dem Umstrukturierungsknoten lässt sich erfahren, ob ein Kunde ein Konto eines bestimmten Typs besitzt und wenn ja, wie hoch der Saldo jedes Kontentyps ist.

Beispieldaten nach der Neustrukturierung

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	Draft	1000	1005.32	1005.32	$null$
12702	Savings	100	144.51	$null$	144.51
12703	Savings	300	321.20	$null$	321.20
12703	Savings	150	204.51	$null$	204.51
12703	Draft	1200	586.32	586.32	$null$

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Beispiel: Umstrukturierungs- & Aggregat-Knoten

Beispieldaten nach der Neustrukturierung

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	Draft	1000	1005.32	1005.32	$null$
12702	Savings	100	144.51	$null$	144.51
12703	Savings	300	321.20	$null$	321.20
12703	Savings	150	204.51	$null$	204.51
12703	Draft	1200	586.32	586.32	$null$

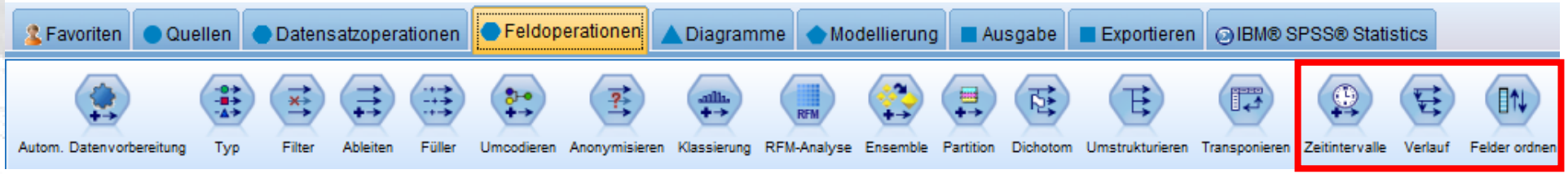
Aggregieren über das Schlüsselfeld CustID mit den Aggregatfeldern Account_Draft_Current_Bal_Sum und Account_Savings_Current_Bal_Sum

Beispieldaten nach der Neustrukturierung und Aggregation

CustID	Record_Count	Account_Draft_Current_ Bal_Sum	Account_Savings_Current_ Bal_Sum
12701	1	1005.32	$null$
12702	1	$null$	144.51
12703	3	586.32	525.71

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Feldoperationen



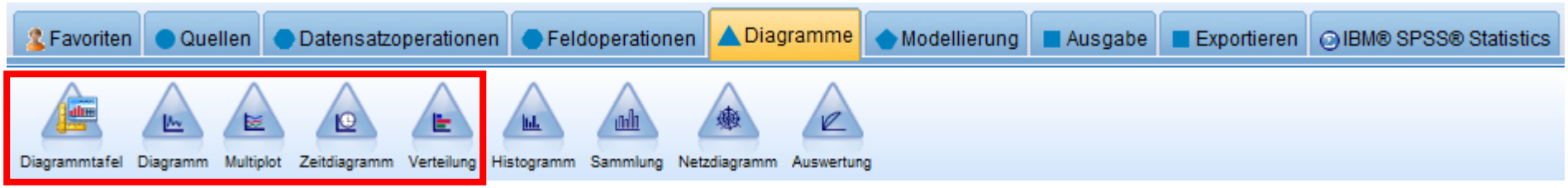
Zeitintervalle: Zur Bestimmung von Intervallen bei Zeitreihendaten (z. B. bei täglich neu generierten Daten)

Verlauf: Generiert neue Datenfelder aus einem bestehenden Feld (z. B. bei Zeitreihen neue Felder für jeden Tag)

Felder ordnen:
Erlaubt die Veränderung der Reihenfolge von Datenfeldern im Stream

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Diagramme



Diagrammtafel: Mehrere Diagrammtypen in einem Knoten

Diagramm: Zeigt die Beziehung zwischen numerischen Feldern mit Hilfe von Linien oder Punkten (Streudiagramm).

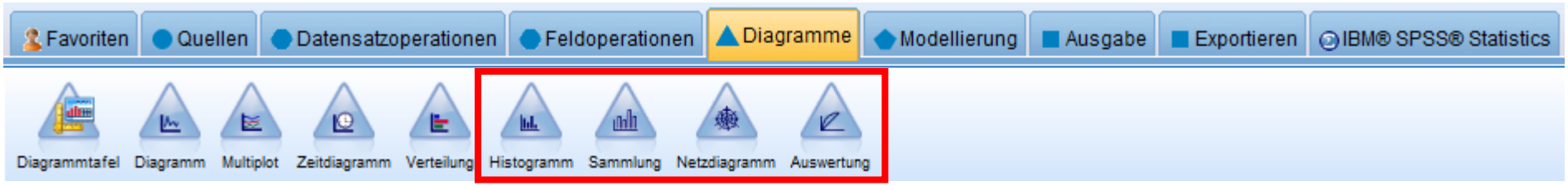
Multiplot: Stellt mehrere Y-Felder über einem einzelnen X-Feld dar. (z. B. Entwicklungen im Zeitverlauf zweier Variablen)

Zeitdiagramm: Stellt Zeitreihen über einen bestimmten Zeitraum dar.

Verteilung: Stellt die Verteilung nicht-numerischer Variablen grafisch dar

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Diagramme



Histogramm: Stellt die Verteilung numerischer Variablen grafisch dar.

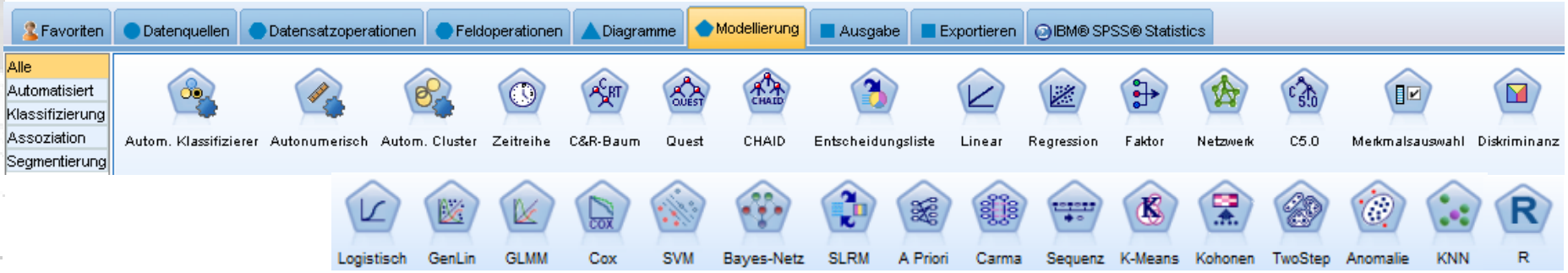
Sammlung: Stellt die Verteilung eines numerischen Feldes relativ zu den Werten eines anderen Feldes dar.

Netzdiagramm: Zeigt die Stärke der Beziehung zwischen den Werten aus mindestens zwei kategorialen Feldern. Die Verbindungsstärke wird durch unterschiedliche breite Linien dargestellt.

Auswertung: Zur grafischen Evaluierung und zum Vergleich von Modellen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Modelle



Netzwerk ***Kohonen*** — Neuronales Netz

C5.0 ***C&R-Baum*** ***Quest*** ***CHAID*** ***SLRM*** ***Entscheidungsliste*** — Entscheidungsbaum

K-Means ***TwoStep*** ***Anomalie*** ***KNN (k-nearest neighbours)*** — Clusteranalyse

A Priori ***Carma*** ***Sequenz*** — Assoziationsanalyse

Linear ***Regression*** ***Logistisch*** ***GenLin*** ***GLMM*** ***Cox*** ***SVM (Support Vector Machines)*** — Regressionsanalyse

Faktor — Faktorenanalyse

Merkmalsauswahl — Merkmale nach Wichtigkeit auswählen

Diskriminanz — Diskriminanzanalyse

Autom. Klassifizierer ***Autonumerisch*** ***Autom. Cluster*** — Kombinationsmodelle

Bayes-Netz — Bayes'sche Netze

Zeitreihen — Zeitreihenanalyse

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Ausgabe

Tabelle: Darstellung der verfügbaren Daten in einer Tabelle

Matrix: Darstellung in einer Kreuztabelle

Analyse: Analyse von Modellen durch verschiedene Vergleiche zwischen den vorhergesagten Werten und den tatsächlichen Werten bei Modellen mit überwachtem Lernen (ermöglicht auch Fehlklassifizierungstabellen)

Data Audit: Ermöglicht einen umfassenden ersten Blick auf die Daten zu werfen (Übersichtsstatistiken, Histogramme, Verteilungsdiagramme, Informationen zu Ausreißern, Extremwerten und fehlenden Werten)

Transform.: Schnelle Sichtprüfung der Verteilung einzelner Datenfelder und Vorschläge zur Transformation. (Ableitungs- oder Füllerknoten lassen sich ggf. generieren, die dann Transformationen durchführen)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Ausgabe

Statistik: Grundlegende Übersichtsdaten zu numerischen Feldern (Mittelwert, Min, Max, Standardabweichung etc.)

Mittelwerte: Vergleich der Mittelwerte zwischen unabhängigen Gruppen oder zwischen Paaren von in Bezug stehenden Feldern (Test auf Signifikanz)

Bericht: Erstellt formatierte Berichte

Globalwerte: Berechnet für ausgewählte Felder globale Werte, wie z. B. Maximum, Minimum, Mittelwert. Nach der Berechnung können die Werte in CLEM-Ausdrücken verwendet werden.
(Bsp: @GLOBAL_MAX(Feldname), @GLOBAL_MEAN(Feldname))

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

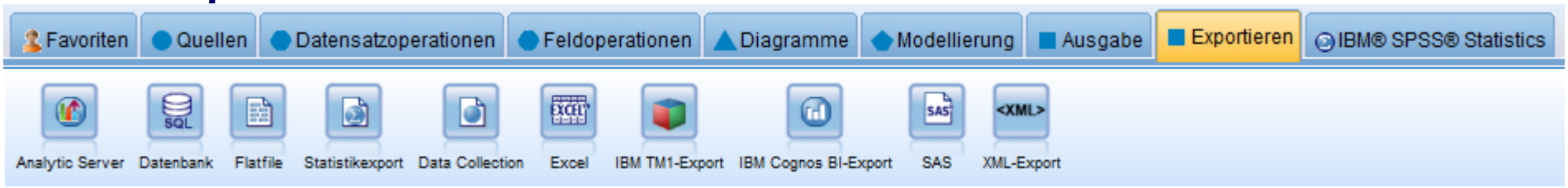
Simulationsanpassung: Erstellung eines Quellenknotens für die Erzeugung künstlicher Daten

R-Ausgabe: Datenanalyse mit R

Simulationsevaluierung: Erstellung eines Berichtes über die Verteilung und Variabilität eines Zielfeldes

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Exportieren



Analytic Server: Export an den Analytic Server

Datenbank: Ausgabe in eine ODBC-kompatible relationale Datenquelle

Flatfile: Ausgabe als Textdatei mit Trennzeichen

Statistikexport: Ausgabe als PASW / SPSS-Datei im .sav-Format

Data Collection: Ausgabe für Marktforschungssoftware SPSS Dimensions

Excel: Ausgabe als Excel-Datei, Achtung: Zeilenbegrenzung!

IBM TM1-Export und Cognos BI-Export:
Export auf den TM1- bzw. Cognos BI-Server

SAS: Ausgabe als Datei im SAS-Format

XML-Export: Ausgabe als XML-Datei

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

IBM SPSS Statistics

Zur Ergänzung von IBM® SPSS® Modeler und seinen Data-Mining-Funktionen bietet Ihnen IBM® SPSS® Statistics die Möglichkeit, weiterführende statistische Analysen durchzuführen und Daten zu verwalten.

Wenn Sie eine lizenzierte Kopie von SPSS Statistics installiert haben, können Sie von SPSS Modeler eine Verbindung aufbauen und komplexe, aus mehreren Schritten bestehende Datenänderungen und Analysen ausführen, die anderenfalls von SPSS Modeler nicht unterstützt werden. Für den erfahrenen Benutzer gibt es auch die Option, die Analysen mithilfe von Befehlssyntax weiter anzupassen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

1. Analytisches CRM
2. IBM SPSS Modeler
3. Knoten-Übersicht
4. Einlesen von Daten
5. Datenaufbereitung
6. Data Mining-Modelle

Skalen

- Daten bilden die Ausgangsbasis jeder Data Mining-Analyse. Sie entstehen durch Messvorgänge.
- Beim Messen werden Eigenschaften von Objekten nach bestimmten Regeln in Zahlen ausgedrückt. Die dabei angewandte „Messlatte“ nennt sich **Skala**.
- Objekt-Eigenschaften können mit unterschiedlichen Skalen „gemessen“ werden. Man unterscheidet verschiedene **Skalenniveaus**.
- Je höher das Skalenniveau, desto größer ist auch der Informationsgehalt der betreffenden Daten und desto mehr Rechenoperationen und statistische Maße lassen sich auf die Daten anwenden.
- **Leider entsprechen die Modeler-Datentypen nicht exakt der gängigen Skalenniveau-Einteilung.** Vor der Anwendung einzelner Funktionen bzw. vor der Interpretation ist daher immer eine inhaltliche Überprüfung der einzelnen Variablen notwendig!

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Skalen-Niveaus im Modeler

Skala		Merkmale	Mögliche rechnerische Handhabung	Beispiel	Modeler
nicht-metrische/ kategoriale Skalen	**Nominalskala**	Klassifizierung qualitativer Eigenschafts-ausprägungen	Bildung von Häufigkeiten	männlich/ weiblich	Nominal Flag
	Ordinalskala	Rangwerte mit Ordinalzahlen	Median, Quartile	Schulnoten	Ordinal
Metrische/ kardinale Skalen	**Intervallskala**	Skala mit gleichgroßen Abschnitten ohne Nullpunkt	Subtraktion, Mittelwert	Zeitskala (Datum)	Stetig
	Verhältnisskala (Ratio-Skala)	Skala mit gleichgroßen Abschnitten mit Nullpunkt	Addition, Division, Multiplikation	Einkommen, Preis, Alter	Stetig

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Speichertypen und Datentypen

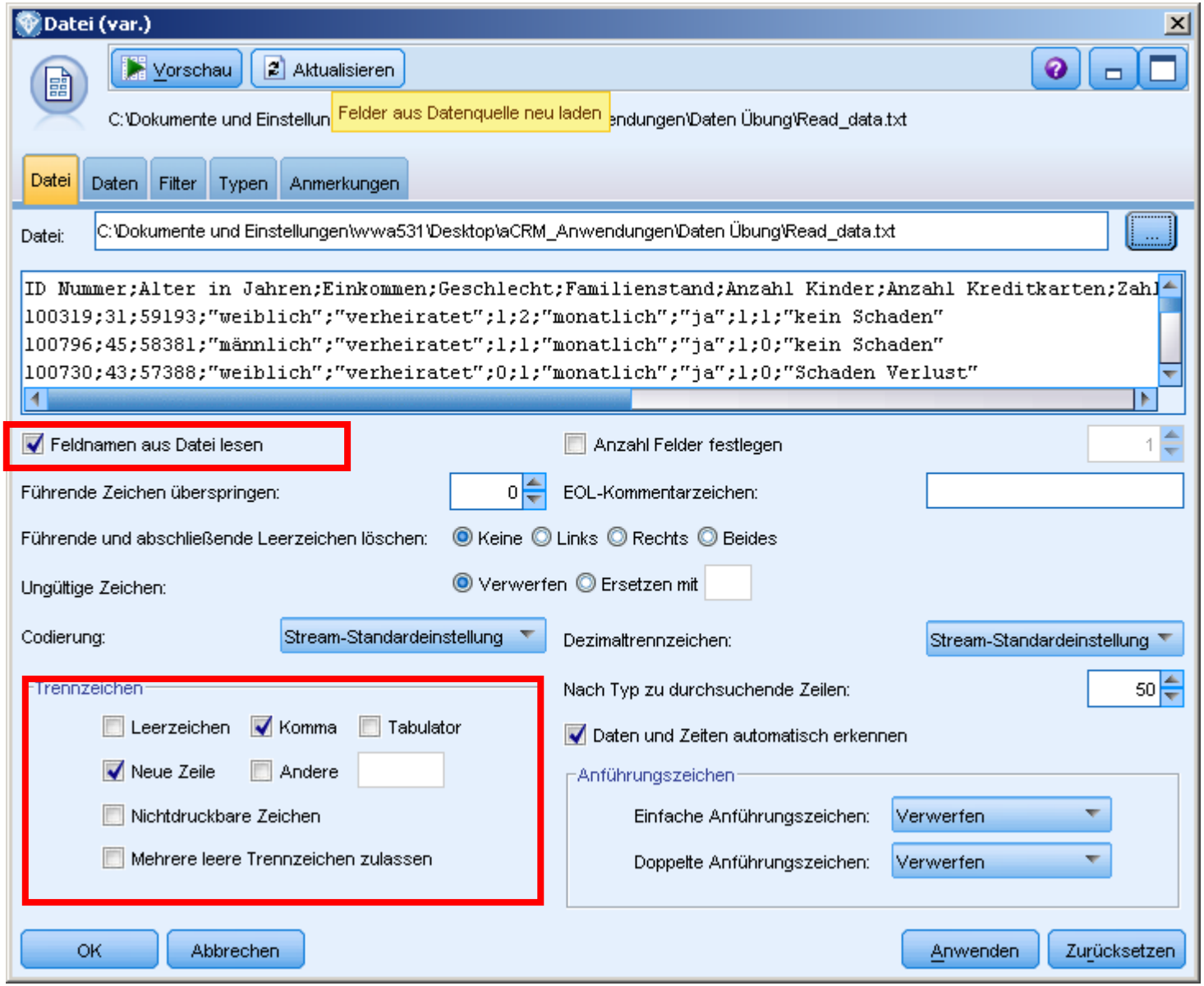
- Grundsätzlich ist zwischen **Speichertypen** und **Datentypen** („Messniveau“) zu unterscheiden!
- Speichertypen werden im **Datenquellen-Knoten** eingestellt, Datentypen können auch mit einem **Typ-Knoten** verändert werden.

Datei (var.): Datenimport aus Textdateien mit Feldern variabler Größe

Typ: „Verwalten“ der Felder
(Festlegen des Skalenniveaus eines Felds, des Wertebereichs, der Missing Values und ob das Feld bei einem maschinellem Lernprozess ein Eingabefeld (Prädikatorfeld) oder ein Ausgabefeld (vorhergesagtes Feld) darstellt)

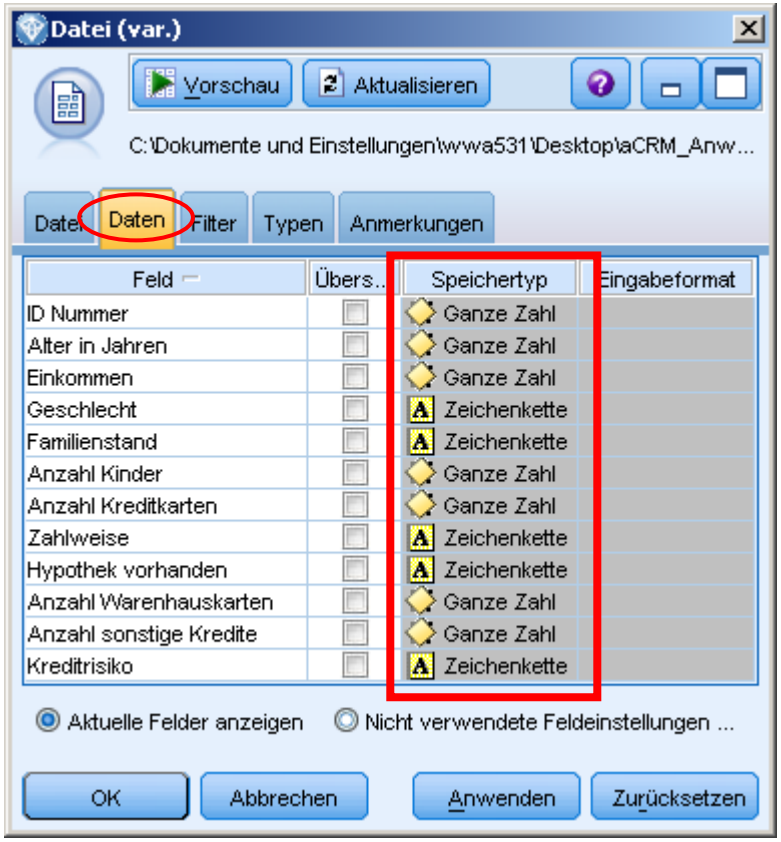
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

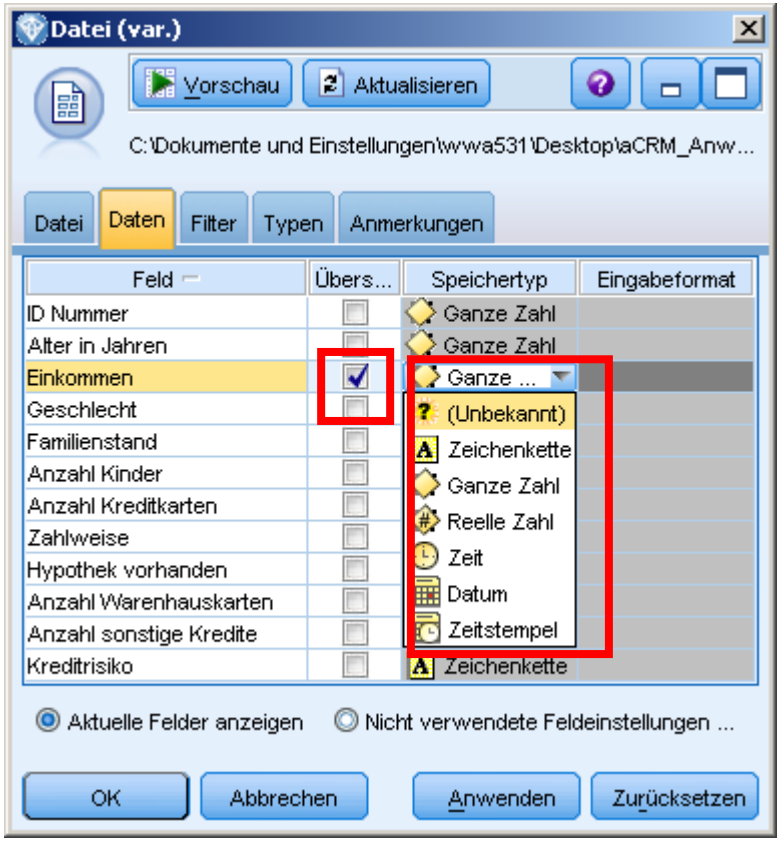
Der Import-Knoten „Datei (var.)“



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Speichertypen



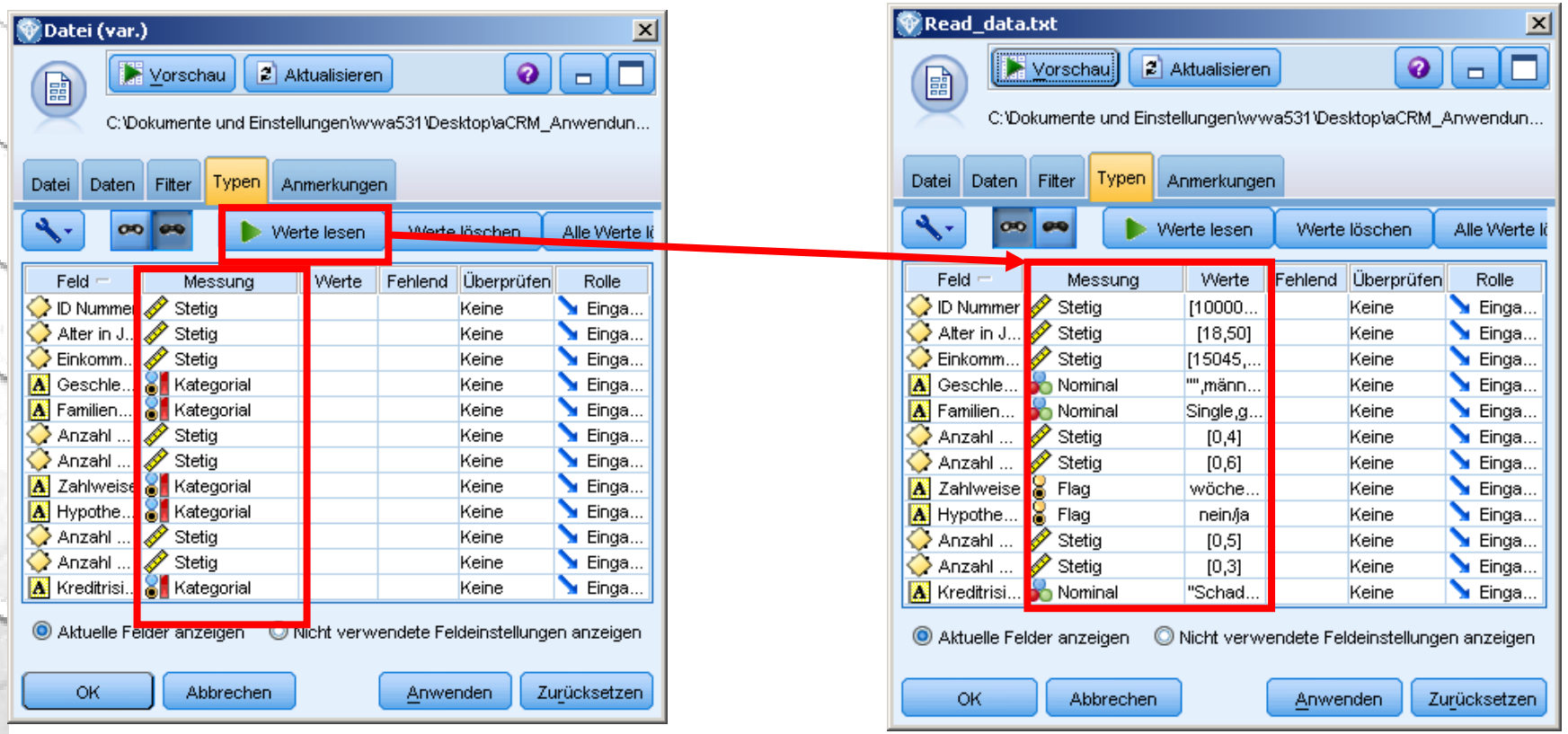


Durch Aktivieren der Check-Box „Überschreiben“ können Speichertypen geändert und Zusatzangaben zum Format (Bsp. Datum) gemacht werden.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datentypen

Der Modeler definiert Felder mit numerischen Inhalten automatisch als **Stetig** und Zeichenketten-Felder automatisch als **Kategorial**. Durch einen Klick auf „Werte lesen“ bzw. beim ersten Durchlauf eines Typ-Knotens werden die Daten abgearbeitet und der vermeintlich richtige Typ eingefügt.



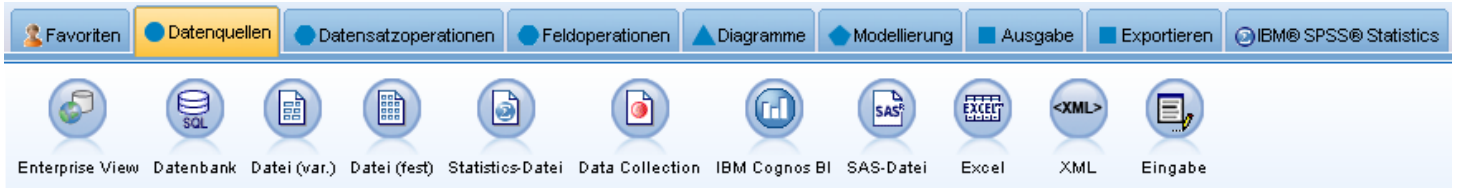
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datentypen im Modeler

Bezeichnung im Modeler	Bedeutung	Beispiel
Stetig	**Numerisch** (ganze Zahlen, reelle Zahlen oder Datum/Uhrzeit)	0 - 100 0,75 - 1,25
Kategorial	**Zeichenkettenwerte**, wenn eine exakte Anzahl unterschiedlicher Werte nicht bekannt ist	ID_ABCD1234
Flag	**Binär**	1/0 Ja/Nein
Nominal	**Nominal**	gelb/grün/blau 1/2/3
Ordinal	**Ordinal**	Mehrere unterschiedliche Werte, die eine natürliche Reihenfolge aufweisen
Ohne Typ	**Keine Zuordnung möglich**	Zahl der Ausprägungen übersteigt die maximale Größe eines Sets

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

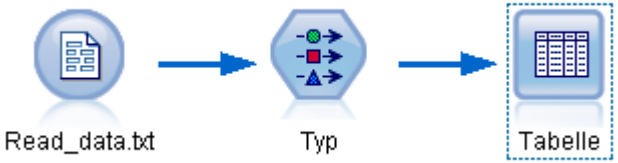
Übung 4.1



Lesen Sie die Datei Read_data.txt ein, und stellen Sie deren Inhalt in einer Tabelle dar.

Welche Veränderung schlagen Sie hinsichtlich der erkannten Datentypen vor?

Zu verwendende Knoten für den Stream:



Speichern Sie den Stream als „Übung 4_1 “

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 4.2 (1/3)

Ausgangssituation: Ein Mobilfunkanbieter möchte Kundenabwanderungen analysieren, um abwanderungsgefährdete Kunden zu erkennen und Gegenmaßnahmen einzuleiten.

Vorüberlegungen zur Data Mining Aufgabe:

1. Wieso/wann kann ein (kostenintensives) Kündigungspräventionsprogramm für das Unternehmen sinnvoll sein?
2. Wie könnten die Rahmenbedingungen für ein solches Data Mining Projekt grundsätzlich aussehen?
3. Welche Daten werden benötigt?
4. Woher können diese Daten stammen?

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 4.2 (2/3)

Aufgabe:

- Lesen Sie die Daten aus der Datei rawdata.txt ein.
- Stellen Sie die Daten mit Hilfe eines Tabelle-Knoten dar.
- Untersuchen Sie die Daten hinsichtlich der zuvor getroffenen Vorüberlegungen zu Analysezweck und Datenbedarf. Ist die Datengrundlage für eine Kündigeranalyse geeignet?
- Speichern Sie den Stream als „Übung 4_2“

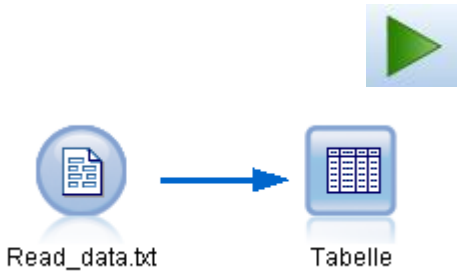


Tabelle (16 Felder, 1.508 Datensätze)

Datei Bearbeiten Generieren

Tabelle Anmerkungen

	ID	LONGDIST	International	LOCAL	DROPPED	PAY_MTHD	LocalBillType	LongDistanceBillType	AGE	SEX	STATUS	CHILDREN	Est_Income	Car_Owner	No phone lines	CHURNED
1	0	5.246	7.515	86.328	0	CH	FreeLocal	Standard	57	F	M	2	27535.3	Y	$null$	Vol
2	3	0.000	0.000	3.942	0	CC	Budget	Intnl_discount	50	F	S	2	64632.3	N	$null$	InVol
3	4	5.556	0.000	9.363	1	CC	Budget	Intnl_discount	68	F	M	2	81000.9	N	$null$	Vol
4	8	14.019	5.680	29.806	0	CC	Budget	Standard	34	M	S	0	87467.1	Y	$null$	Current
5	9	0.000	0.000	3.327	0	CC	FreeLocal	Standard	37	F	M	0	67988.2	Y	$null$	
6	10	13.664	2.956	32.638	0	CC	FreeLocal	Intnl_discount	60	M	M	2	83220.6	N	1	Vol
7	11	0.000	0.000	1.413	0	CC	FreeLocal	Standard	84	F	S	0	$null$	N	$null$	InVol
8	13	0.281	0.000	8.537	0	CH	Budget	Intnl_discount	28	F	M	2	20850.4	N	$null$	Vol
9	17	1.577	0.000	19.981	0	CC	FreeLocal	Standard	52	M	S	0	84112.6	N	$null$	Current
10	19	11.031	0.000	34.278	0	CC	Budget	Standard	87	F	S	2	3776.12	N	$null$	Vol
11	20	0.453	0.000	73.012	0	Auto	FreeLocal	Standard	88	F	M	2	73865.9	Y	$null$	Vol
12	22	3.729	0.000	18.847	0	Auto	FreeLocal	Standard	76	M	M	2	30933.6	Y	$null$	Current
13	23	20.295	0.000	76.052	0	CC	FreeLocal	Standard	76	F	M	1	12309.6	N	$null$	Vol
14	26	10.370	2.153	24.668	0	CH	FreeLocal	Standard	87	M	S	0	69864.0	N	$null$	Vol
15	28	20.268	0.000	102.864	0	CH	Budget	Intnl_discount	90	F	M	0	91620.6	N	$null$	Current
16	30	25.528	0.747	5.186	0	CC	Budget	Standard	62	M	S	0	96501.9	Y	$null$	Current
17	31	26.495	0.000	31.085	0	CC	FreeLocal	Intnl_discount	81	M	S	1	3968.54	N	$null$	Current
18	32	8.082	0.000	10.586	0	CH	Budget	Standard	50	F	M	2	13774.2	N	$null$	Vol
19	33	2.946	0.261	77.238	0	CH	FreeLocal	Standard	48	M	S	2	39428.2	N	$null$	Current
20	40	24.346	0.000	62.826	0	CC	Budget	Standard	37	F	S	2	4988.14	N	$null$	Current
21	42	8.865	4.437	43.644	0	CH	Budget	Intnl_discount	55	M	S	2	85753.8	N	$null$	Vol
22	49	26.862	0.000	15.492	0	CC	FreeLocal	Intnl_discount	21	M	M	2	23564.1	N	$null$	Vol
23	55	17.208	8.942	138.045	0	CC	FreeLocal	Intnl_discount	42	M	S	0	37661.0	N	$null$	Current

OK

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 4.2 (3/3)

Beschreibung der Datenfelder

Datenfelder	Bedeutung	Datensatz 1
ID	Kundennummer	0
LONGDIST	Zeit für Ferngespräche pro Monat	6.246
International	Zeit für Internat. Gespr. p Monat	7.515
LOCAL	Zeit für Ortsgespräche p. Monat	86.328
DROPPED	Anzahl abgebrochene Gespräche	0
PAY_MTHD	Zahlungsweise	CH
LocalBillType	Tarifmodell Ortsgespräche	FreeLocal
LongDistanceBillType	Tarifmodell Ferngespräche	Standard
AGE	Alter	57
SEX	Geschlecht	F
STATUS	Familienstand	M
CHILDREN	Anzahl Kinder	2
Est_Income	Geschätztes Einkommen	27536.3
Car_Owner	KfZ-Besitzer	Y
CHURNED	Current – aktueller Kunde Vol – freilwilliger Kündiger InVol – unfreiwilliger Kündiger	Vol

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

1. Analytisches CRM
2. IBM SPSS Modeler
3. Knoten-Übersicht
4. Einlesen von Daten
5. Datenaufbereitung
6. Data Mining-Modelle

Ziel

- Zum Erstellen und Anwenden von Modellen benötigt man eine saubere Datentabelle. Dazu werden folgende Schritte ausgeführt:
  - Zusammenführen unterschiedlicher Datenquellen
  - Bilden neuer, aussagekräftiger Variablen
  - Aggregation auf die betrachtete Analyseeinheit (Granularität)
    Bsp: Bankkunde – Separate Datensätze für jedes Konto, für jede Person, oder für jeden Haushalt?
- Bei der Datenaufbereitung stellen sich die Fragen:
  - Sind die Modell-Anforderungen erfüllt?
    (z. B. Skalenniveaus, Datentyp)
  - Ist ein hohes Datenqualitätsniveau gewährleistet?
    (z. B. Missing Values)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenanalyse

- Einen schnellen, groben Überblick über die Daten erhält man mit dem Data Audit-Ausgabe-Browser (durch **Ausführen des Data Audit-Knotens**)

Data Audit von [15 Felder]

Datei Bearbeiten Generieren

Audit Qualität Anmerkungen

Feld	Diagramm	Messung	Min	Max	Mittelwert	Std.Abw.	Schiefe	Kategorien	Fälle gültig
ID		Stetig	0	4999	2508.382	1468.658	-0.019	--	1508
LONGDIST		Stetig	0.000	29.982	13.672	9.401	0.011	--	1508
International		Stetig	0.000	9.951	0.829	2.223	2.699	--	1508
LOCAL		Stetig	0.018	450.624	51.595	54.814	2.288	--	1508
DROPPED		Stetig	0	4	0.121	0.421	4.381	--	1508
PAY_MTHD		Nominal	--	--	--	--	--	3	1508
LocalBillType		Flag	--	--	--	--	--	2	1508
LongDistanceBillType		Flag	--	--	--	--	--	2	1508

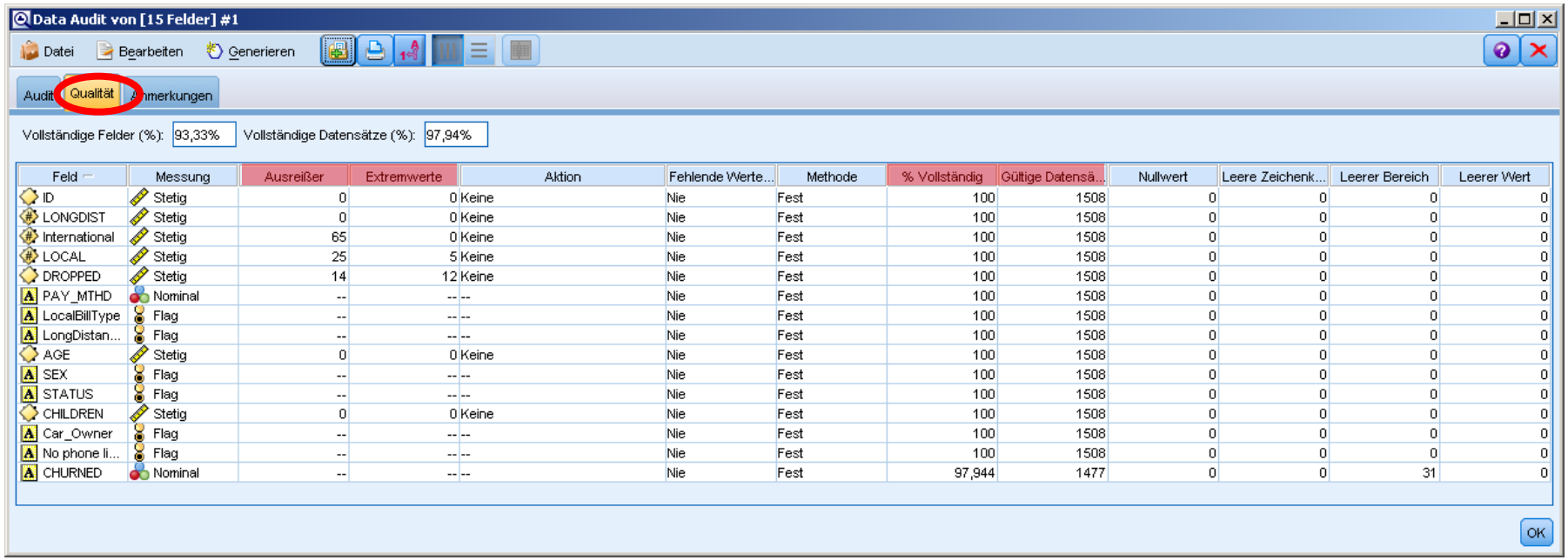
¹ Zeigt ein Mehrfachmodus-Ergebnis an ² Zeigt ein Stichprobenergebnis an

OK

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenqualität

- Ausreißer, Extremwerte und fehlende Werte erkennt man durch Klick auf die **Registerkarte „Qualität“** im Data Audit-Ausgabe-Browser:



Feld	Messung	Ausreißer	Extremwerte	Aktion	Fehlende Werte...	Methode	% Vollständig	Gültige Datensä...	Nullwert	Leere Zeichenk...	Leerer Bereich	Leerer Wert
ID	Stetig	0	0	Keine	Nie	Fest	100	1508	0	0	0	0
LONGDIST	Stetig	0	0	Keine	Nie	Fest	100	1508	0	0	0	0
International	Stetig	65	0	Keine	Nie	Fest	100	1508	0	0	0	0
LOCAL	Stetig	25	5	Keine	Nie	Fest	100	1508	0	0	0	0
DROPPED	Stetig	14	12	Keine	Nie	Fest	100	1508	0	0	0	0
PAY_MTHD	Nominal	--	--	--	Nie	Fest	100	1508	0	0	0	0
LocalBillType	Flag	--	--	--	Nie	Fest	100	1508	0	0	0	0
LongDistan...	Flag	--	--	--	Nie	Fest	100	1508	0	0	0	0
AGE	Stetig	0	0	Keine	Nie	Fest	100	1508	0	0	0	0
SEX	Flag	--	--	--	Nie	Fest	100	1508	0	0	0	0
STATUS	Flag	--	--	--	Nie	Fest	100	1508	0	0	0	0
CHILDREN	Stetig	0	0	Keine	Nie	Fest	100	1508	0	0	0	0
Car_Owner	Flag	--	--	--	Nie	Fest	100	1508	0	0	0	0
No phone li...	Flag	--	--	--	Nie	Fest	100	1508	0	0	0	0
CHURNED	Nominal	--	--	--	Nie	Fest	97,944	1477	0	0	31	0

Ungültige Werte

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Behandlung von Ausreißern

- Oftmals gibt es bei Merkmalen Ausreißer nach oben und nach unten: Ausprägungen mit unerwartet hohen oder niedrigen Werten.
- Behandlungsansätze:
  - **Keine Änderungen vornehmen**
    Einige Data Mining-Verfahren sind gegenüber Ausreißern robust.
  - **Datensätze herausfiltern**
    Kann die Analyse verbessern, jedoch auch zu einer Verzerrung der Daten führen.
  - **Spalte ignorieren**
    Möglicherweise kann das betreffende Merkmal durch andere treffendere Informationen ersetzt werden.
  - **Ausreißer-Werte ersetzen**
    Mögliche Ersatzwerte: „null“, Durchschnittswert, andere - als sinnvoller erachtete – Maximum-/Minimum-Werte oder aus anderen Merkmalen abgeleitete Werte.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Quelle: Berry/Linoff (2000): Mastering Data Mining, S. 160 f.

Übung 5.1 (1/3)

- Die Datei Holiday_all.txt enthält Daten eines Reiseanbieters.
  - Kontrollieren und korrigieren Sie ggf. die Speicher- und Datentypen aller Datenfelder
  - Ordnen Sie die Datenfelder wie folgt:

Feldname	Beschreibung
CUSTID	Kundennummer
NAME	Kundenname
DOB	Geburtsdatum Kunde
GENDER	Geschlecht
REGION	Heimatort
NUMPARTY	Anzahl Reiseteilnehmer
HOLCOST	Kosten der Reise
NIGHTS	Übernachtungen
TRAVDATE	Reisestart
HOLCODE	Reise-Kennung
COUNTRY	Land
POOL	Schwimmbad
ACCOM	Typ der Unterkunft
DIST_TO_BEACH	Distanz zum Strand

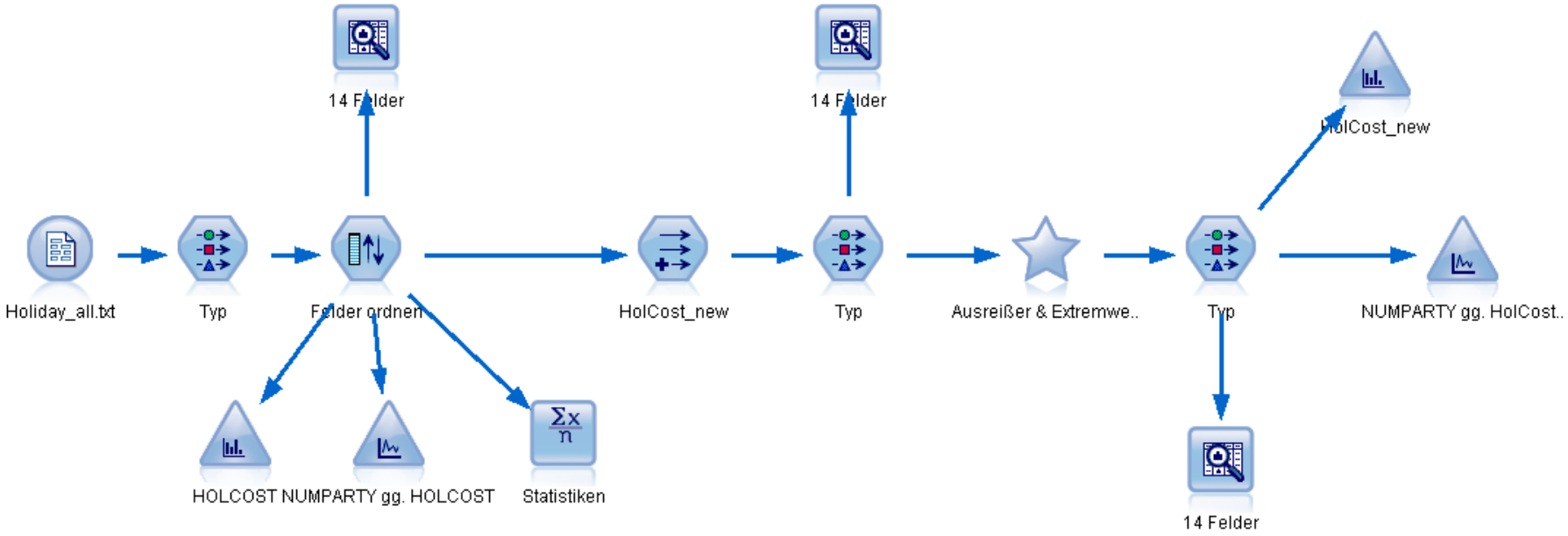
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.1 (2/3)

- Untersuchen Sie Qualität der Daten.
  Wie würden Sie hier mit Ausreißern umgehen?
- Untersuchen Sie das Feld **HOLCOST** anhand eines Histogramms (Klassenbreite 10,0).
  Wie könnte man bzgl. dieses Feldes mit Ausreißern umgehen?
- Untersuchen Sie die Beziehung der Variablen **HOLCOST** und **NUMPARTY** anhand eines Plots und einer statistischen Analyse.
  Welche Aussage können Sie treffen?
  Besteht eine Korrelation zwischen den beiden Datenfeldern?
- Leiten Sie das neue Feld **HolCost_new** als Kopie des Feldes **HOLCOST** ab. Fügen Sie einen Typ-Knoten zur Initialisierung des neuen Feldes an, bevor Sie einen Data Audit-Knoten anhängen und ausführen.
- Erzwingen Sie für das Feld **HolCost_new** eine Aktion (Hier: Angleichen der Ausreißer auf einen best. Maximalwert). Generieren Sie dann einen Superknoten für Ausreißer & Extremwerte für dieses Feld.
  Welche Aktionen werden durchgeführt?
- Verbinden Sie den Superknoten mit dem Stream.
- Untersuchen Sie das Feld **HolCost_new** anhand eines Histogramms und dessen Beziehung zum Feld **NUMPARTY** durch einen Plot.
  Welche Veränderungen hat sich zum ursprünglichen Feld ergeben?
- Speichern Sie den Stream unter „Übung 5_1“

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.1 (3/3)



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Missing Values

- Fehlende Werte werden in Datenbank meist durch den Wert „**NULL**“ ersetzt
- **Arten fehlender Werte:**
  - **Leere Datenzellen**
    Beispielsweise bleibt das Datenfeld „Telefonnummer“ leer, falls ein Kunde seine Nummer nicht mitteilen will, weil er sich vor ungewünschten Anrufen fürchtet. Doch selbst in diesem Unterschied steckt eine Information, die gespeichert werden kann.
  - **Nicht existierende Werte**
    Zum Beispiel kann der „Umsatz pro Jahr“ erst nach dem Ablauf eines Jahres ermittelt werden.
  - **Unvollständige Daten**
    Treten z. B. bei der Anreicherung um externe Marketing-Daten auf, falls einzelne Datensätze nicht zur Struktur der externen Daten passen.
  - **Nicht gesammelte Daten**
    Informationen, die nicht gespeichert wurden.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Quelle: Berry/Linoff (2000): Mastering Data Mining, S. 177 ff.

Missing Values

- Nullwerte (nicht definierte Werte), Leerwerte, leere Bereiche und leere Zeichenketten werden vom Modeler **automatisch** als ungültige Werte behandelt.

Nullwert	System Null ($null$)
Leere Zeichenkette	Leeres Feld (auch keine Leerzeichen)
Leerer Bereich	Keine sichtbaren Zeichen (Leerzeichen, Tab...)
Leerer Wert	Alle vom User im Typ-Knoten definierten Blanks.

- **Benutzerdefinierte fehlende Werte:**

Im Typ-Knoten doppelt auf das entsprechende Datenfeld klicken.
Im sich öffnenden Fenster lassen sich durch Aktivieren des Punkts „Fehlende Werte definieren“ bestimmte Eingaben als „Fehlende Werte“ festlegen.
(Bsp: „unbekannt“, 999)

Im Data Audit-Ausgabe-Browser werden diese Werte als „Leere Werte“ gezählt.

Nullwert	Leere Zeichenkette	Leerer Bereich	Leerer Wert
0	0	0	1

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Maßnahmen bei Missing Values (1/2)

- **Keine Änderungen vornehmen**
  Einige Data Mining-Methoden sind gegenüber fehlenden Werten robust.
- **Datensätze herausfiltern**
  Führt zu einer Verzerrung der Daten, falls fehlende Werte systematisch bzw. in ganz bestimmten Fällen auftreten. Die verbleibenden Datensätze wären dann nicht repräsentativ für die Grundgesamtheit. Häufig sind zu viele Datensätze betroffen.
  Diese Maßnahme wird vor allem dann gewählt, wenn die entsprechenden Datensätze in mehreren Datenfeldern Missing Values vorweisen.
- **Datenfelder ignorieren**
  Falls einige Datenfelder viele Missing Values enthalten, ist es oftmals sinnvoll, sich auf die vollständigen Merkmale zu konzentrieren.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Quelle: Berry/Linoff (2000): Mastering Data Mining, S. 177 ff.

Maßnahmen bei Missing Values (2/2)

- **Imputation**
  - Fehlende Werte können prognostiziert werden.
  - Weniger ausgeklügelt, aber in begründeten Fällen dennoch sinnvoll ist es, fehlende Werte durch Konstanten zu ersetzen - zum Beispiel durch: Mittelwert (numerische Daten), Median (ordinale Daten) oder durch den häufigsten Wert (nominale Daten).
- **Separate Modelle bilden**
  Häufig und gerade im Fall nicht-existierender Daten ist es möglich, das Problem zu umgehen, indem man die Kunden basierend auf den vorhandenen Daten segmentiert.
- **Änderung operationaler Systeme**
  Beispielsweise kann durch die Änderung von Dateneingabemasken erreicht werden, dass bislang nicht-gespeicherte Informationen zukünftig erfasst werden. Dieser Ansatz ist allerdings rein in die Zukunft gerichtet.

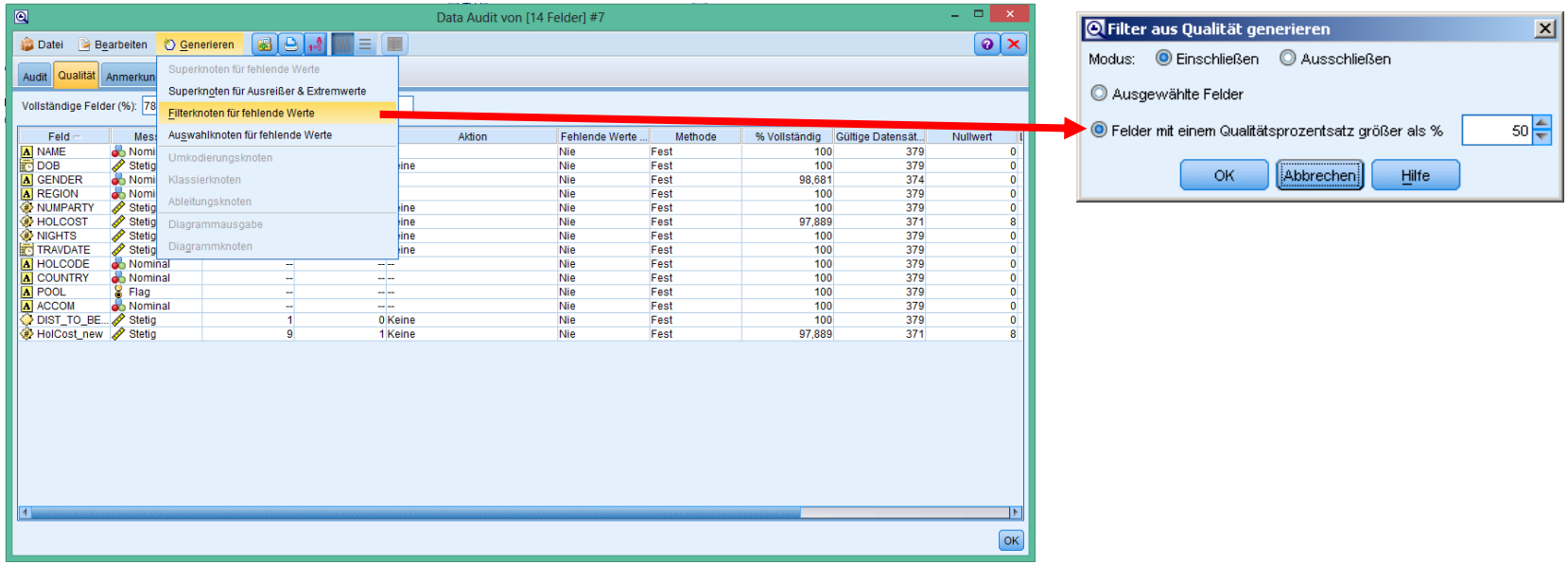
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Quelle: Berry/Linoff (2000): Mastering Data Mining, S. 177 ff.

Behandlung fehlender Werte

Ausschluss von Feldern mit zu vielen fehlenden Werten:

- Mit dem **Data Audit-Knoten** können Sie Felder nach Qualität filtern.

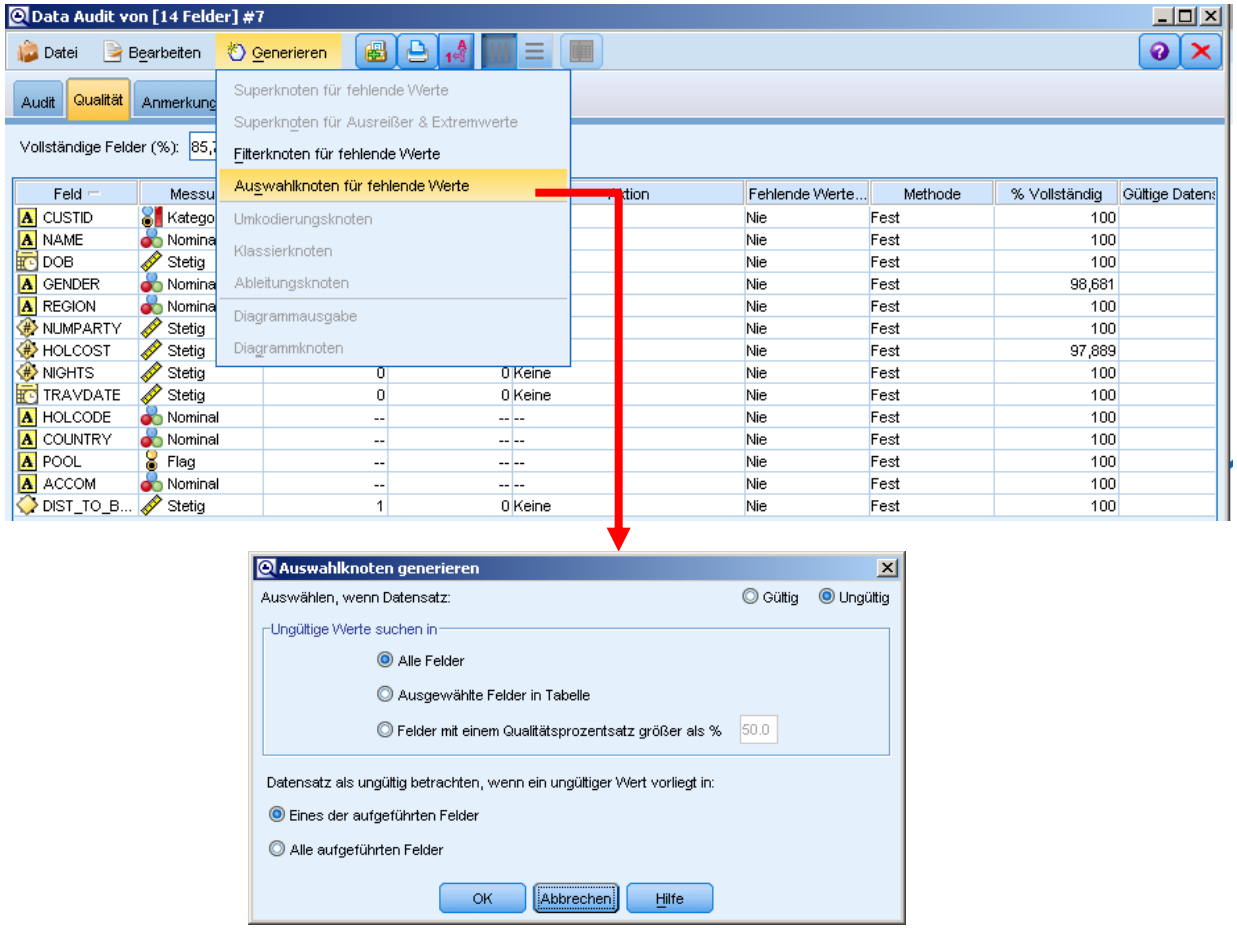


Statt die Felder zu entfernen, können Sie den Wert des Feldes im Bereich „Verwendung“ eines **Typknotens** auf „Keine“ festlegen. Dadurch werden die Felder im Daten-Set belassen, aber aus den Modellierungsprozessen ausgeschlossen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Behandlung fehlender Werte

- **Datensätze mit fehlenden Werten zur weiteren Analyse und Bearbeitung auswählen:**
  - Aus dem **Data Audit-Knoten** einen Auswahlknoten für fehlende Werte generieren.



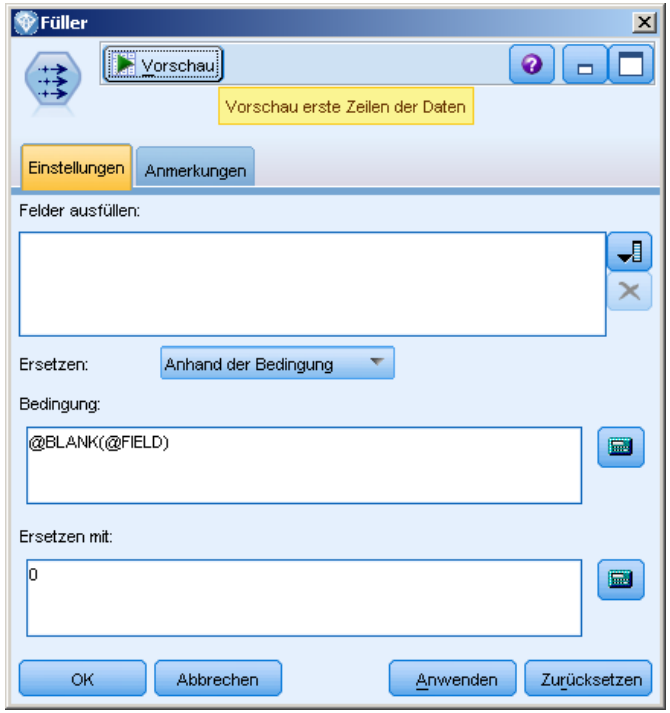
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Behandlung fehlender Werte

Leerfelder auffüllen:

Mit einem Füller-Knoten können Leerfelder mit gültigen Informationen gefüllt werden.

@BLANK(@FIELD) → Alle Datensätze, deren Werte gemäß den Regeln zum Umgang mit Leerstellen, die in einem weiter oben im Stream gelegenen Typknoten oder Quellenknoten (Registerkarte "Typen“) festgelegt wurden, Leerstellen sind.



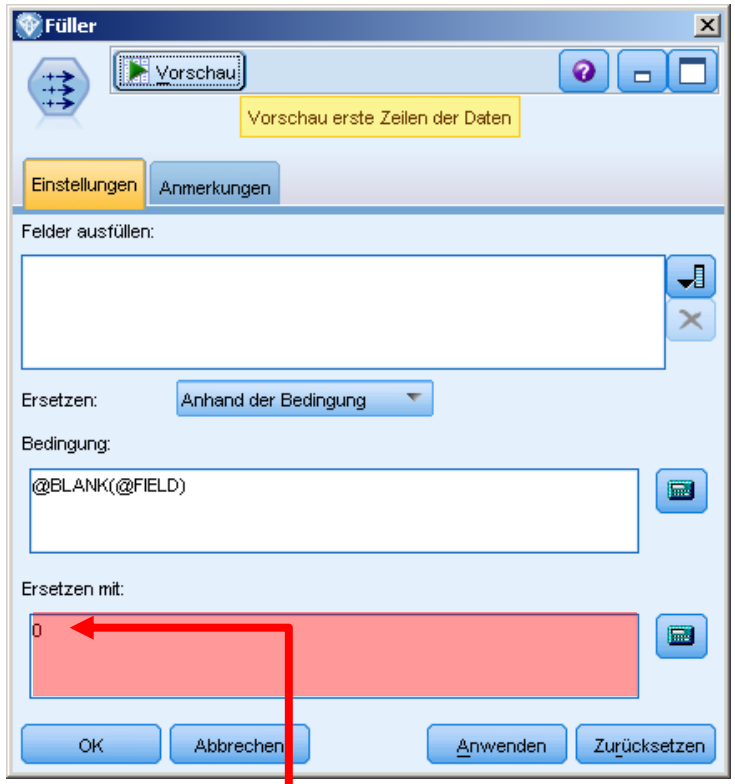
Nur Null-Felder z. B.: @NULL(@FIELD)

Anmerkung: Wenn die ausgewählten Felder den Typ "Zeichenkette" aufweisen, sollten diese mit einem Zeichenkettenwert ersetzt werden. Die Verwendung des Standardwertes 0 oder eines anderen numerischen Wertes als Ersatzwert für Zeichenkettenfelder führt zu einem Fehler.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

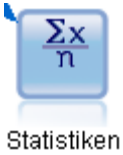
Behandlung fehlender Werte

- **Imputation**:
  - Im Füllerknoten den Standardwert 0 durch gewünschten Wert ersetzen.



Beispiel: Fehlende Werte durch den Mittelwert des entsprechenden Feldes ersetzen.

→ Zunächst mit einem Statistiken-Knoten den Mittelwert berechnen.

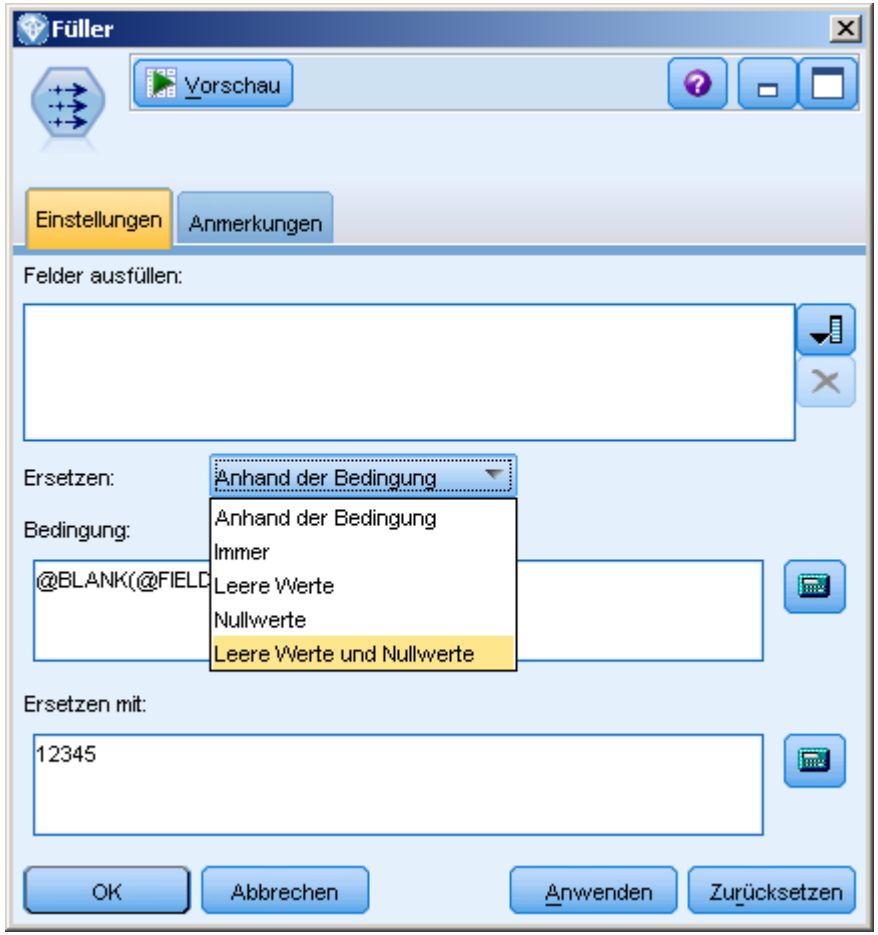


→ Standardwert 0 durch berechneten Mittelwert ersetzen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Behandlung fehlender Werte

- **Imputation**:
  - Alternative durch Auswahl des Drop-Down-Felds „Ersetzen“



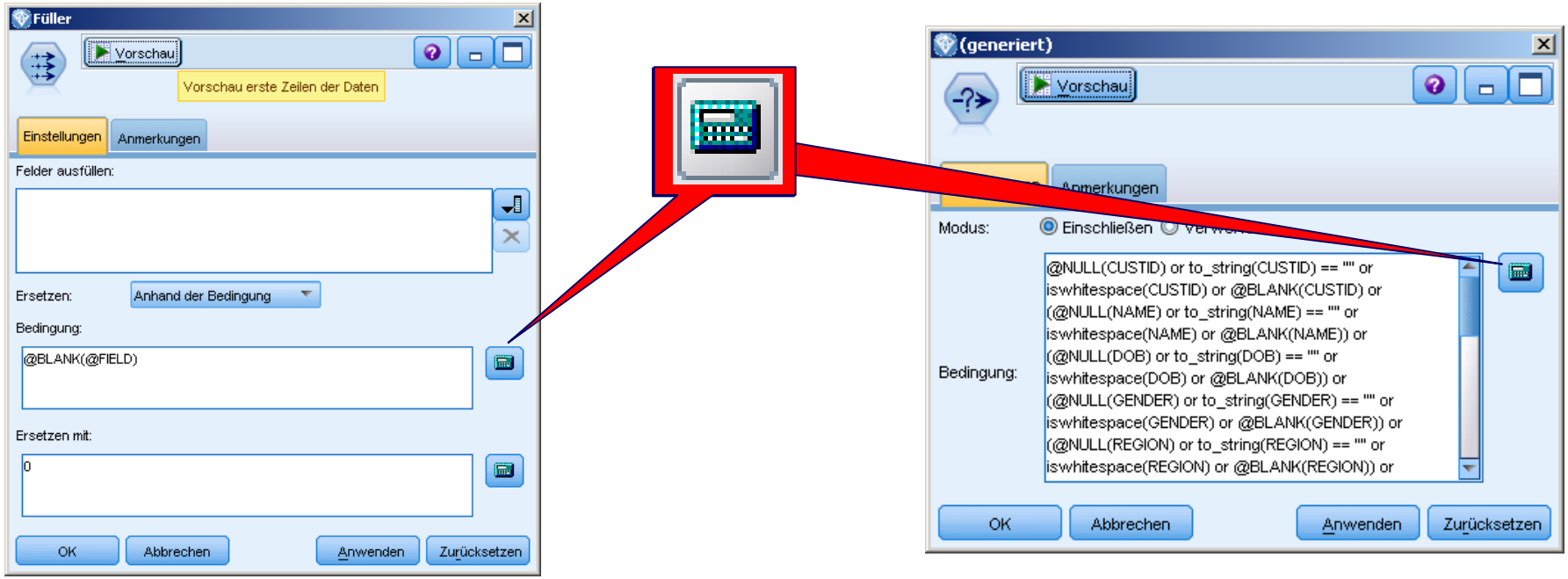
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenauswahl

Filter- und Auswählen-Knoten

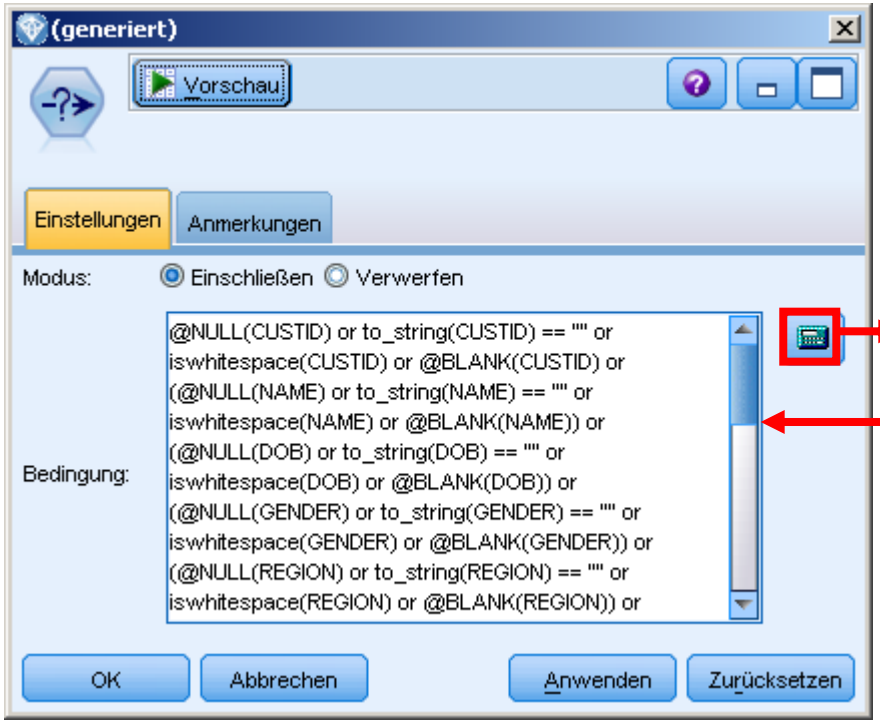
- Anstatt Filter- und/oder Auswählen-Knoten generieren zu lassen (aus dem Data Audit-Knoten heraus), können diese auch direkt in den Stream eingebunden werden.

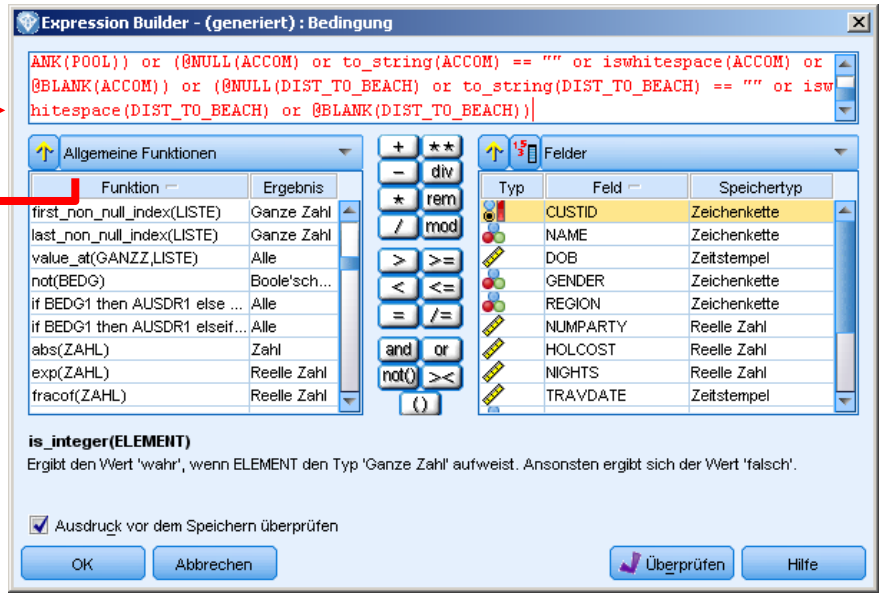
- Zur Definition von Bedingungen etc. steht der „**Expression Builder**“ zur Verfügung.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Expression Builder





Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

CLEM

- Die Ausdrücke können auch in der Sprache **CLEM** (Modeler Language for Expression Manipulation) selbst geschrieben werden. **Beispiele:**
  - issubstring(SUBSTRING,STRING)
    Gibt den Wert 1 zurück, wenn eine bestimmte Zeichenfolge in einer Zeichenkette gefunden wird
    z. B. zur Analyse eines Facebook Nachrichtenfeld auf “WFI”
    → Feld WFI_affiliated = 1
  - If ... Then ... Else
    z. B. zur Schlussfolgerung einer unbekannten Merkmalsausprägung
    IF issubstring(WFI, Nachrichtenfeld) = 1 AND PLZ = “85049“ THEN
    WFI_Student = TRUE
  - time_hours_difference(Zeit1, Zeit2)
    Gibt die Zeitspanne an Stunden aus, die zwischen der Zeitmessung in Feld Zeit1 und in Feld Zeit2 liegt
    z. B. die Websiteverweildauer
    AngryBirdCall = (29.06.2012,10:02:45)
    AngryBirdExit = (29.06.2012,19:45:23)
    Verweildauer = time_hours_difference(AngryBirdCall , AngryBirdExit) → 10 (aufgerundet)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.2 (1/2)

- **Fehlende Werte durch Mittelwert ersetzen**
  (aufbauend auf Stream „Übung 4_2“):
  - Kontrollieren und korrigieren Sie ggf. die Speicher- und Datentypen aller Datenfelder
  - Ermitteln Sie den Mittelwert des Felds „Est_Income“
  - Ersetzen Sie alle fehlenden Werte dieses Felds mit dem Mittelwert.
  - Überprüfen Sie das Ergebnis.
  - Macht dieses Vorgehen in unserem Falle Sinn?
  - Wie könnte man bei der Verwendung eines Mittelwerts noch vorgehen?

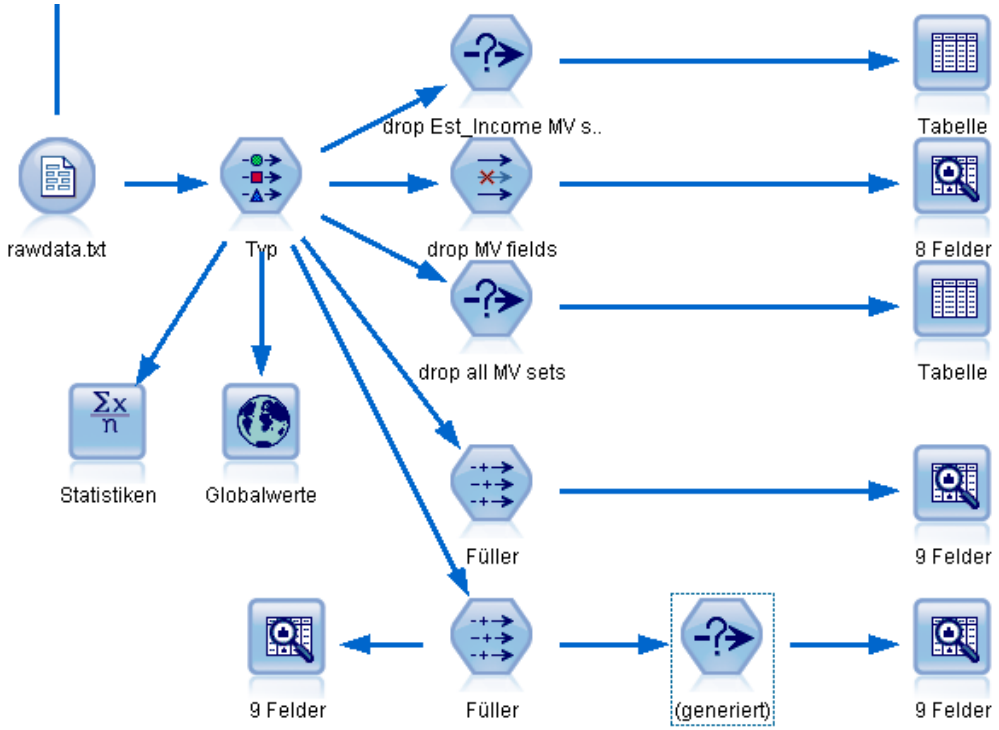
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.2 (2/2)

- Wie könnte eine „optimale“ Lösung zur Behandlung von Missing Values in unserem Beispiel aussehen?
- Führen Sie die Datenbereinigung durch und überprüfen Sie die Ergebnisse.
- Stream speichern unter… „Übung 5_2“

So könnte der Stream aussehen…

... bei unbekannten Knoten kann die Modeler-Hilfe verwendet werden! ☺



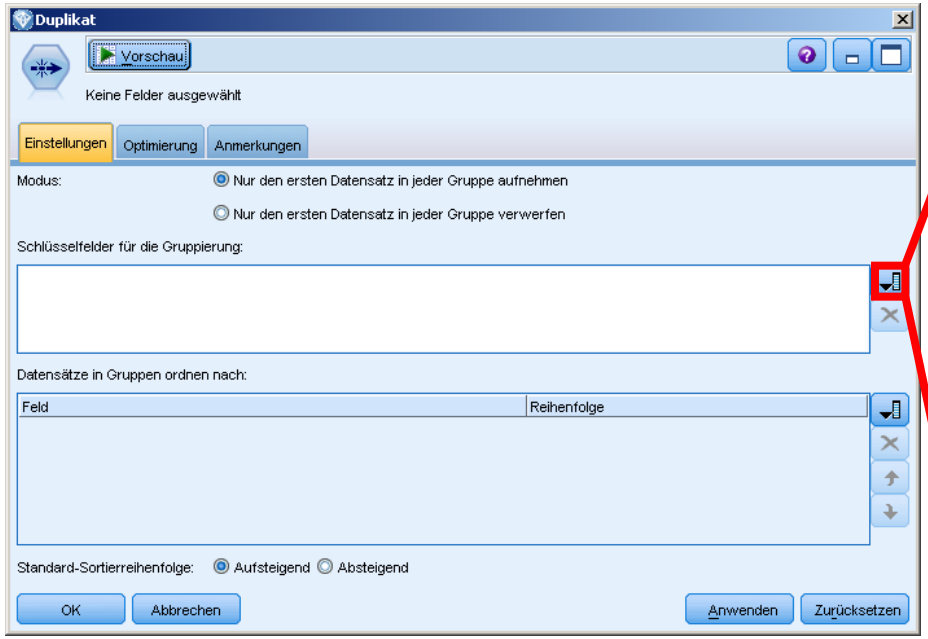
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

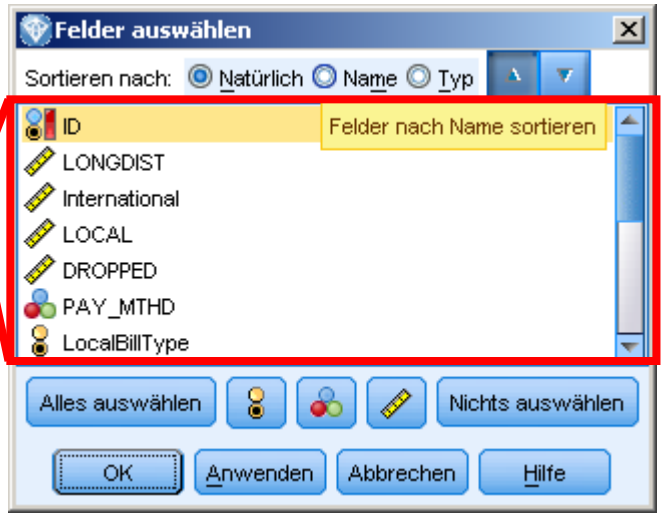
Datenqualität - Duplikate

Duplikat-Knoten

Mit einem **Duplikat-Knoten** können doppelte Datensätze anhand mehrerer Felder erkannt werden.

- Einschließen: Nur jeweils den ersten Datensatz in den Stream aufnehmen.
- Verwerfen: Den ersten gefundenen Datensatz verwerfen und stattdessen etwaige doppelte Datensätze an den Daten-Stream übergeben.
- Zusammengesetzten Datensatz für jede Gruppe erstellen: Aggregation für nichtnumerische Felder.





Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Aggregate

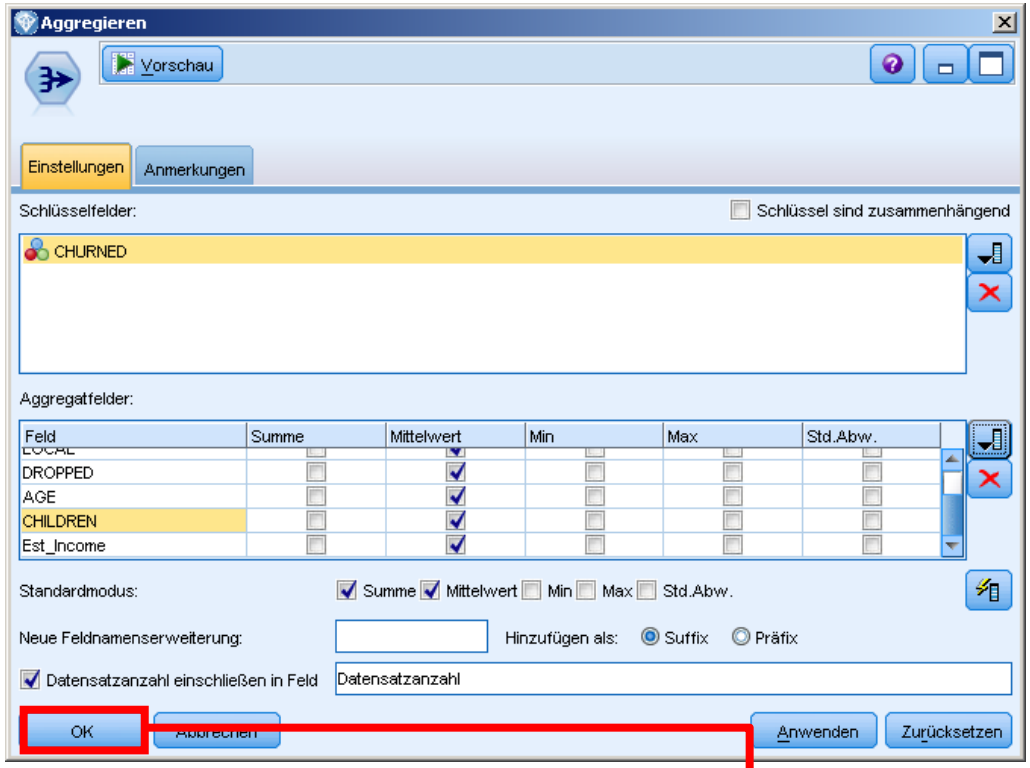
Aggregieren-Knoten

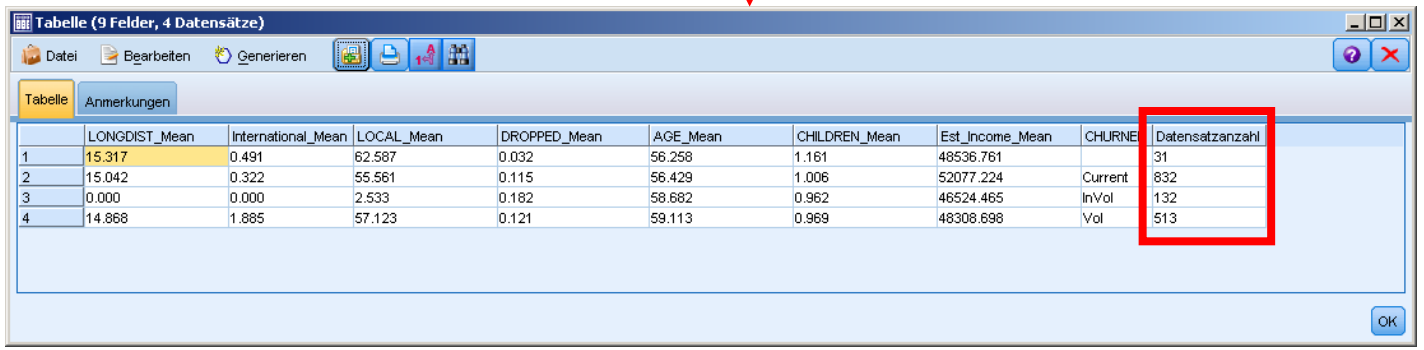
- Zur Datenreduktion können Daten mit dem **Aggregieren-Knoten** zusammengefasst werden.
- Die Ausprägungen der ausgewählten Aggregatfelder werden dann per Aggregatfunktion (z. B. Summe oder Mittelwert) **zu einem Wert je Merkmalsausprägung des Schlüsselfelds verdichtet**.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Aggregate

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

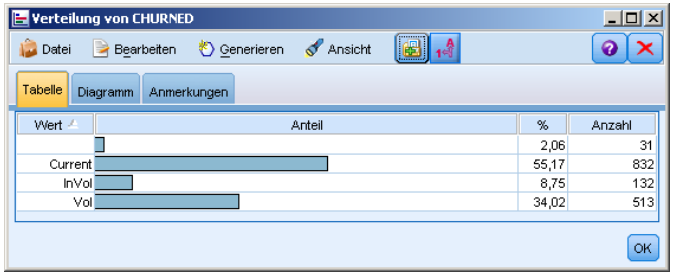




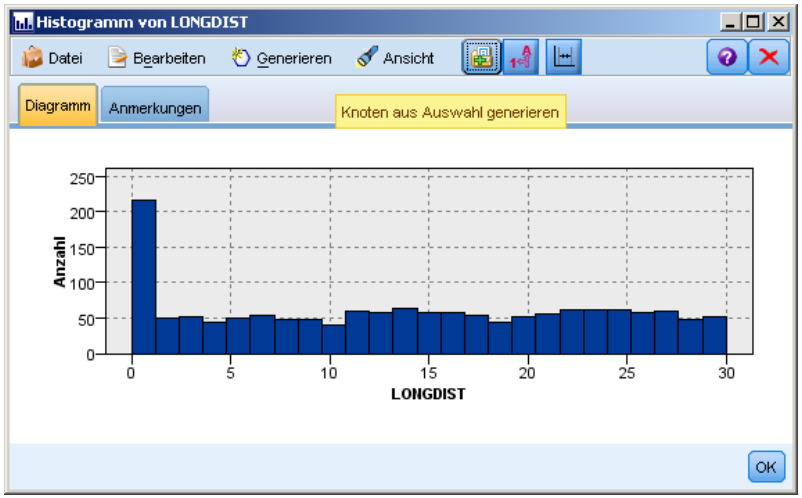
Transformationen

- Bevor Transformationen durchgeführt werden, empfiehlt es sich, im Sinne einer explorativen Datenanalyse, die **Verteilung der Variablen** zu untersuchen.
- Dafür stehen im Bereich Diagramme zwei Knoten zur Verfügung:

Verteilung: Untersuchung nominaler Variablen



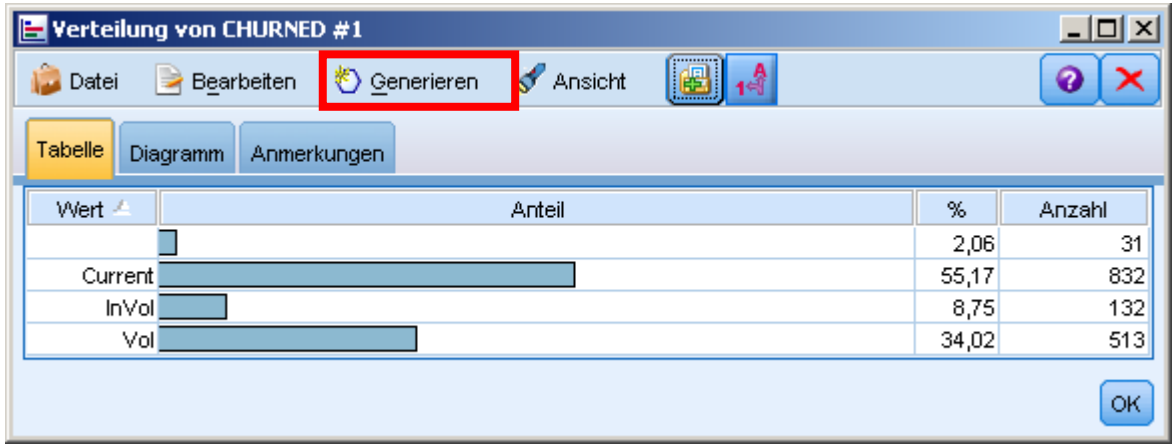
Histogramm: Untersuchung metrischer (bzw. als metrisch skalierte) Variablen



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Transformationen

- Schiefe Verteilungen erschweren das Erstellen korrekter Prognosemodelle.
- **Beispiel:** Ein Neuronales Netz würde bei einer Mailing-Response-Quote von 3% versuchen, jeden Adressat als Nicht-Reagierer zu klassifizieren, um eine Genauigkeit von 97 % zu erreichen.
- Durch Auswahl des Menüpunkts „Generieren” kann direkt ein **Balancierungs-Knoten** erstellt werden, um die ungleiche Verteilung auszugleichen.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

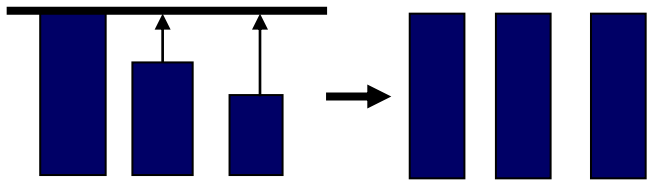
Transformationen

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Balancierungsknoten (erhöhen)

(generiert)

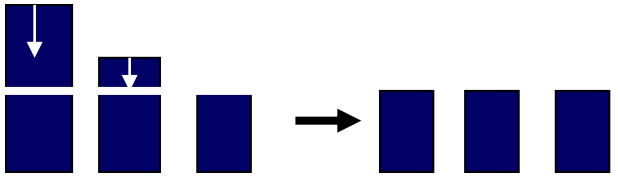
- Balancierungsknoten, der die Größe kleinerer Teilgruppen verstärkt.



Balancierungsknoten (verringern)

(generiert)

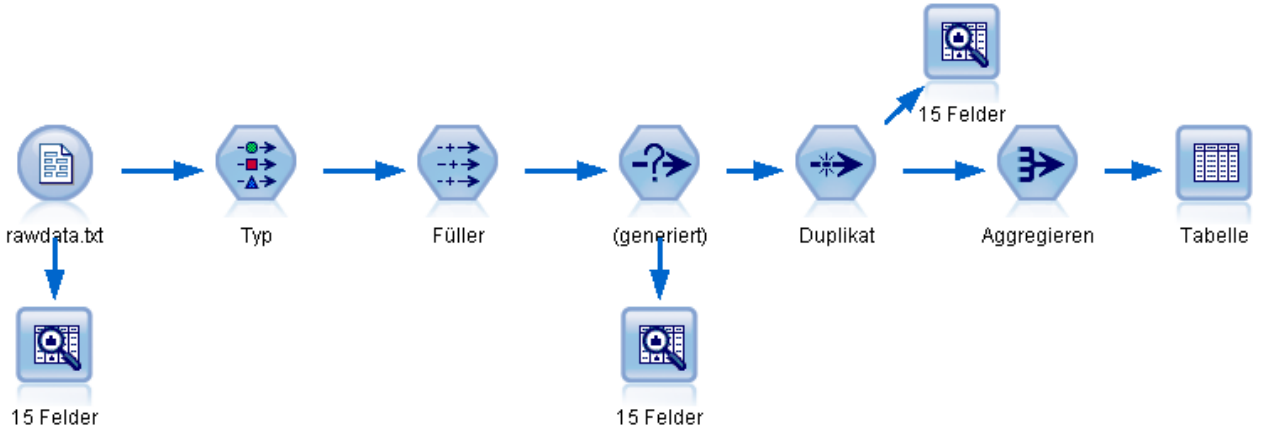
- Balancierungsknoten, der die Größe größerer Teilgruppen verringert.



Übung 5.3

Bitte verwenden Sie den Stream 5_2:

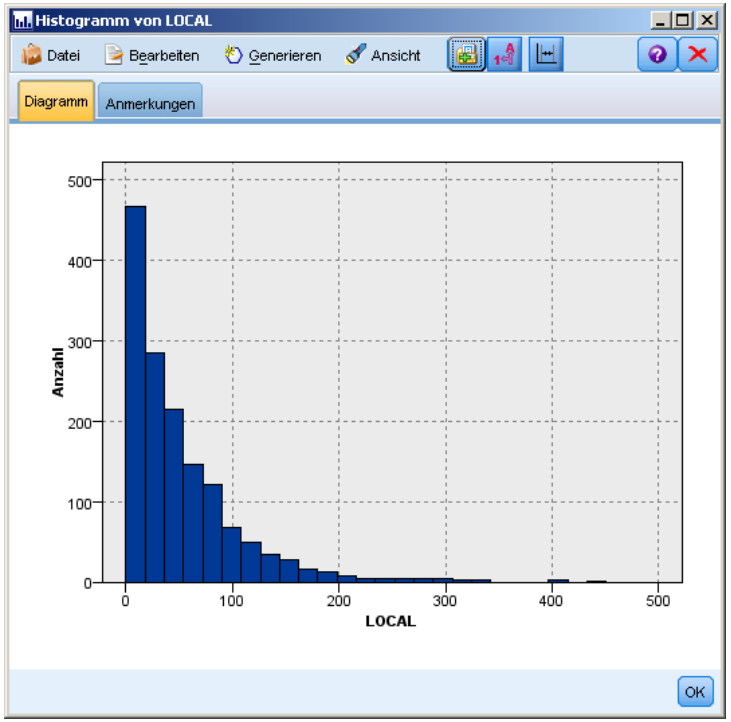
- Welche Maßnahme zur Balancierung der Variable Churned würden Sie empfehlen?
- Fügen Sie dem vorhandenen Stream einen Duplikate-Knoten hinzu. Liegen Duplikate von Kundendatensätzen vor?
- Untersuchen Sie die Mittelwerte der unterschiedlichen Churned-Kategorien bzgl. der Datenfelder Local, LongDist und International über eine Tabelle. Welche Auffälligkeiten stellen Sie fest?
- Stream speichern unter… „Übung 5_3“.

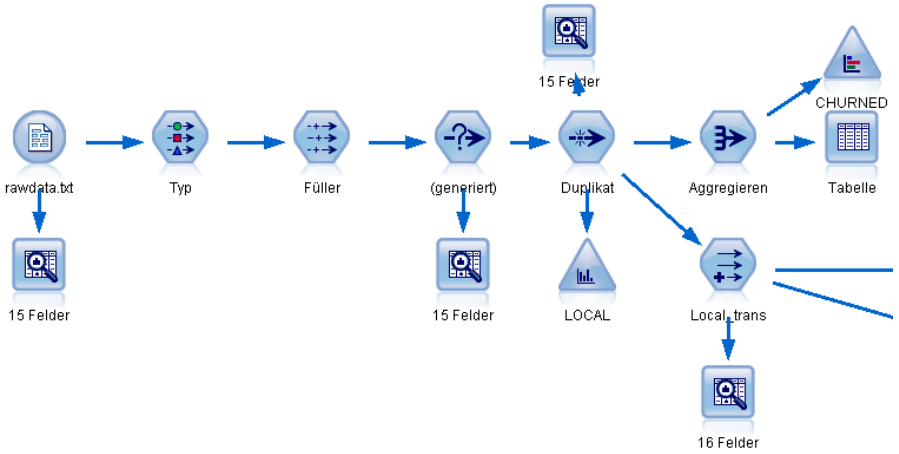


Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.4

- Untersuchen Sie mit einem **Histogramm-Knoten** die Verteilung des Datenfelds „Local“ (Ortsgespräch-Minuten pro Monat).
- Die Verteilung ist sehr schief. Mithilfe einer Logarithmus-Funktion kann die ursprüngliche Schiefe umgekehrt werden. Erstellen Sie dazu über einen Ableiten-Knoten ein neues Feld „Local_trans“ mit der Formel log(LOCAL + 3).
- Stream speichern unter… „Übung 5_4“

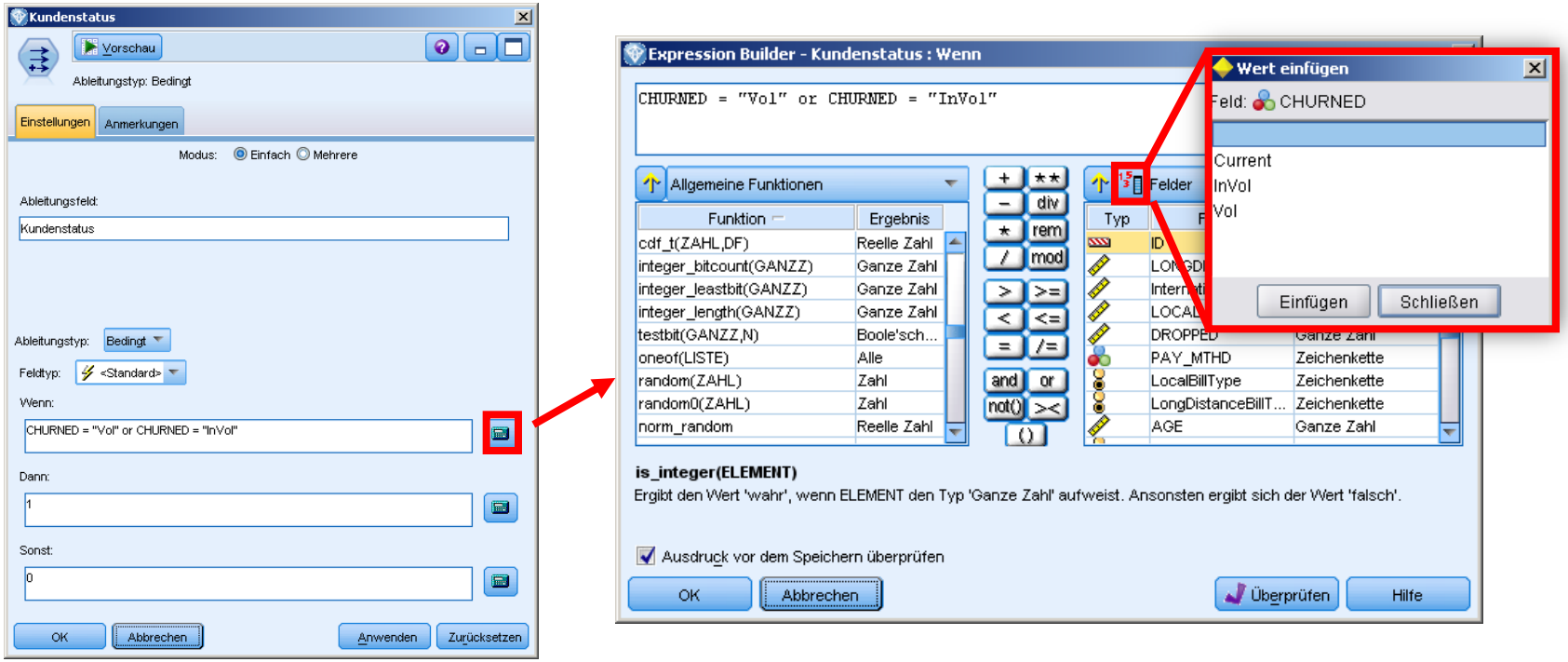




Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Dichotomisieren

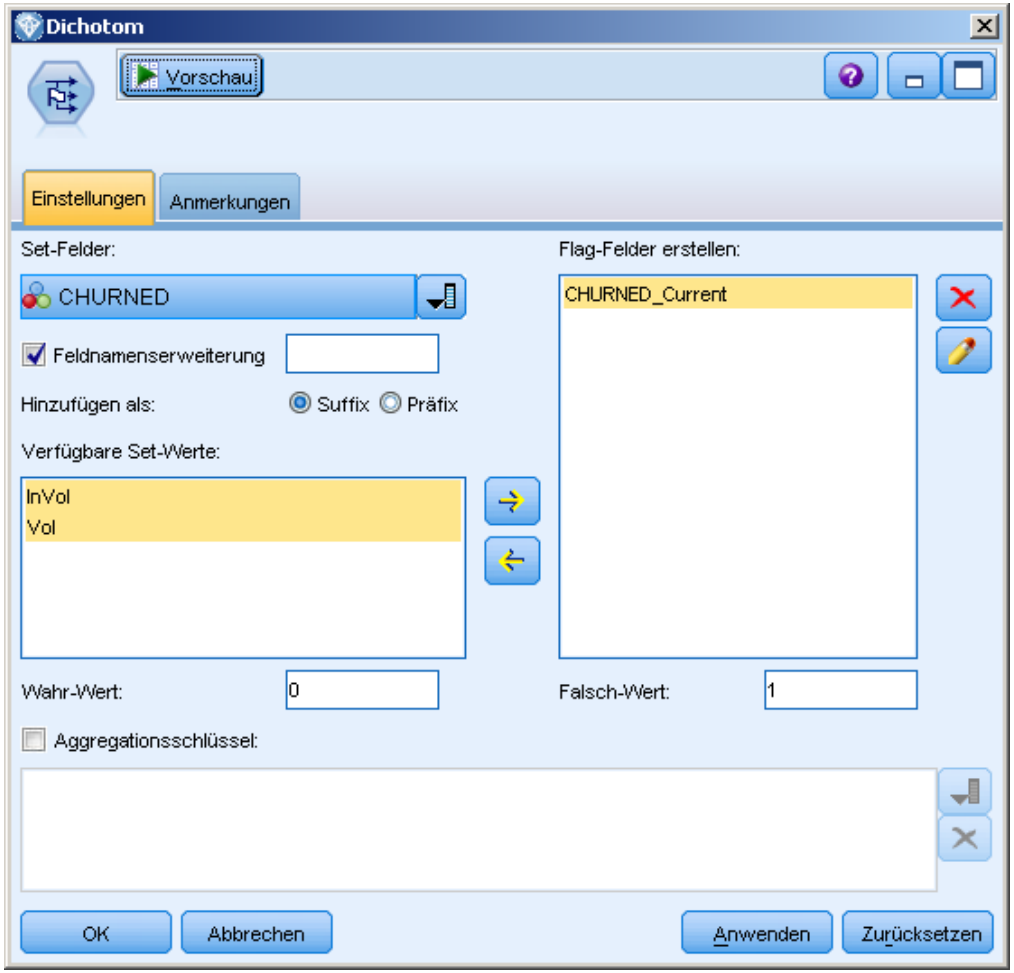
- Eine weitere einfache Transformation ist die Umkodierung in binäre Werte (z. B. 0/1), die ebenfalls durch einen **Ableiten-Knoten** umgesetzt werden kann.
- **Beispiel:**
  Ein neues Datenfeld soll abgeleitet werden, aus dem hervorgeht, ob eine Person aktuell Kunde des Unternehmens ist (CHURNED = „Current“) oder ob sie zum gegenwärtigen Zeitpunkt kein Kunde mehr ist (CHURNED = „Vol“ or CHURNED = „InVol“).



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Dichotomisieren

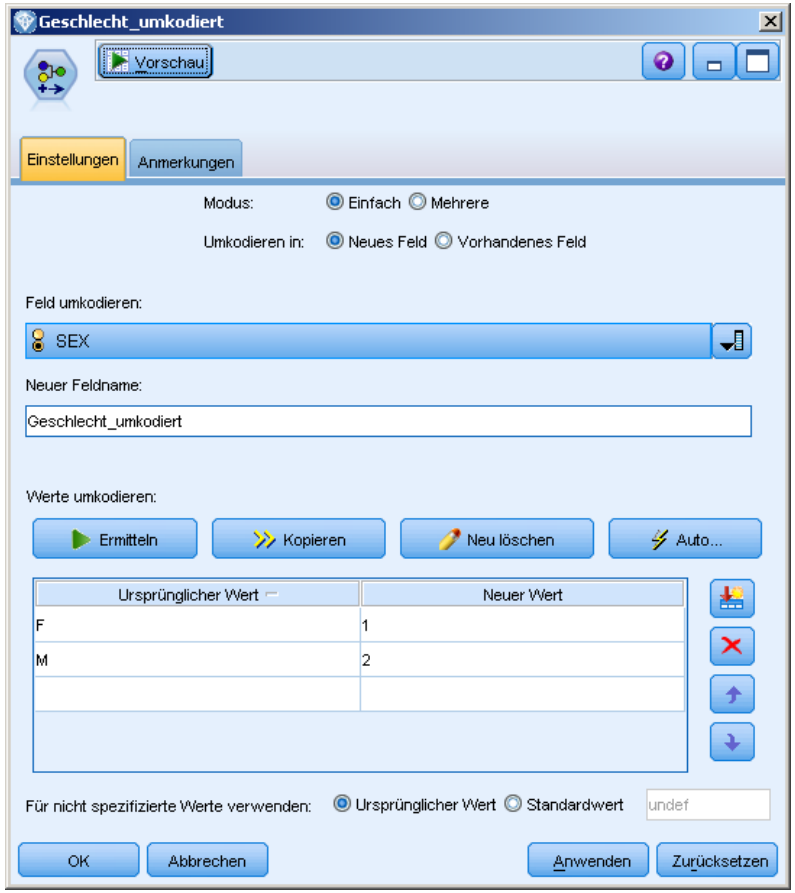
Ein neues binäres Datenfeld kann auch mit einem **Dichotom-Knoten** erstellt werden.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Umkodieren

Eine weitere Form der Transformation ist das Umkodieren vorhandener Werte. Dazu kann im **Umkodieren-Knoten** die Zuordnung der Originalwerte zu neuen Werten festgelegt werden.

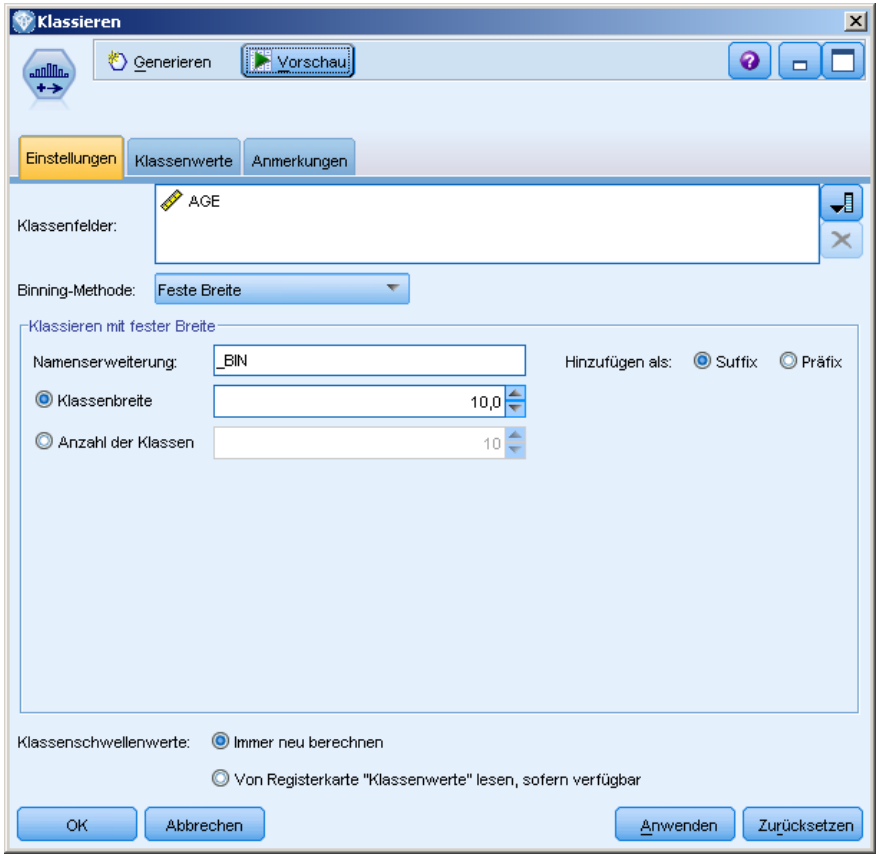


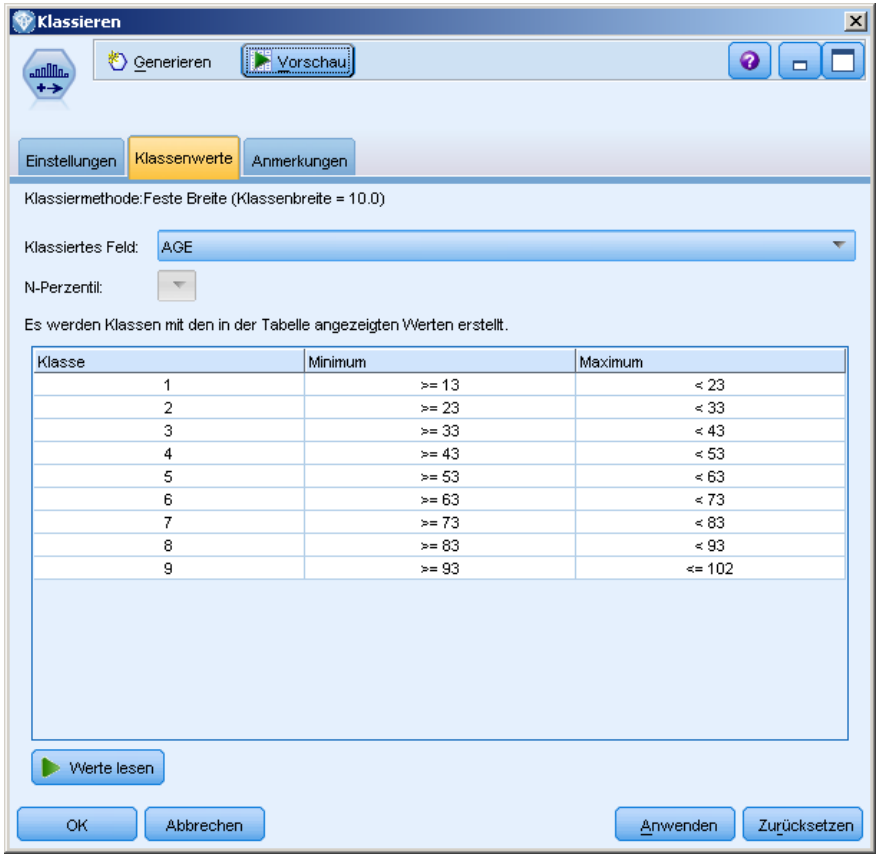
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Klassieren

Mit einem **Klassieren-Knoten** können Kategorien einer Variable gebildet werden (sog. Binning).

- Beispiel: Altersgruppen

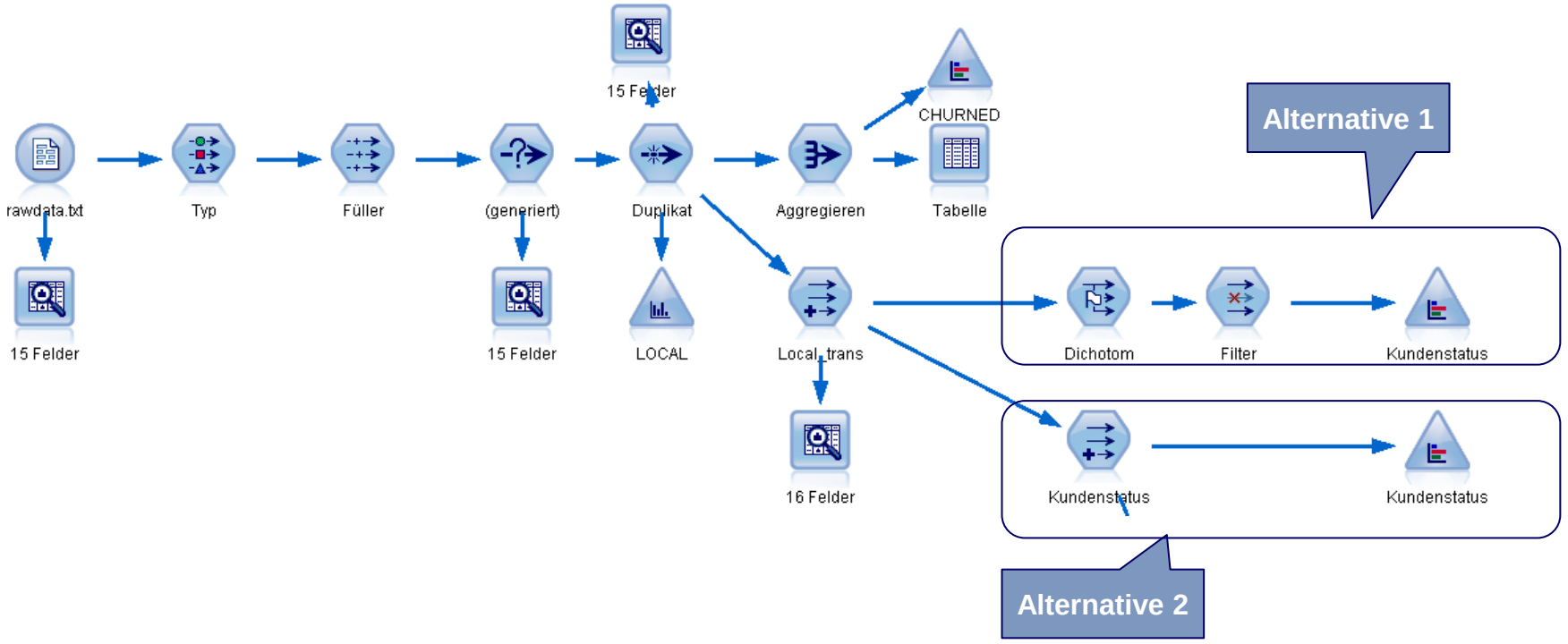




Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.5 (1/2)

- Erzeugen Sie das Feld „Kundenstatus“ mit den oben beschriebenen Informationen und überprüfen Sie die korrekte Zuordnung. Untersuchen Sie anschließend die Verteilung „Kunden“ zu „Nicht-Kunden“.
- Stream speichern unter… „Übung 5_5“



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.5 (2/2)

Outputs

Tabelle (18 Felder, 1.477 Datensätze) #1

Datei Bearbeiten Generieren

Tabelle | Anmerkungen

	TUS	CHILDREN	Est_Income	Car_Owner	No phone lines	CHURNED	Local_trans	Kundenstatus
1		2	27535.300	Y	$null$	Vol	4.492	1
2		2	83220.600	N	1	Vol	3.573	1
3		0	88413.400	Y	$null$	Vol	3.401	1
4		1	34317.600	N	$null$	Vol	4.537	1
5		1	99248.100	N	$null$	Vol	3.812	1
6		2	8749.800	N	$null$	Vol	2.980	1
7		0	81420.000	N	$null$	Vol	3.788	1
8		0	17499.200	N	$null$	Current	2.601	0
9		1	76213.200	N	$null$	Current	3.959	0
10		2	11434.600	N	$null$	Vol	2.716	1
11		2	27362.600	Y	$null$	InVol	1.307	1
12		2	37155.000	Y	$null$	Vol	4.893	1
13		2	75004.500	N	$null$	Current	3.561	0

OK

Verteilung von Kundenstatus #3

Datei Bearbeiten Generieren Ansicht

Tabelle | Diagramm | Anmerkungen

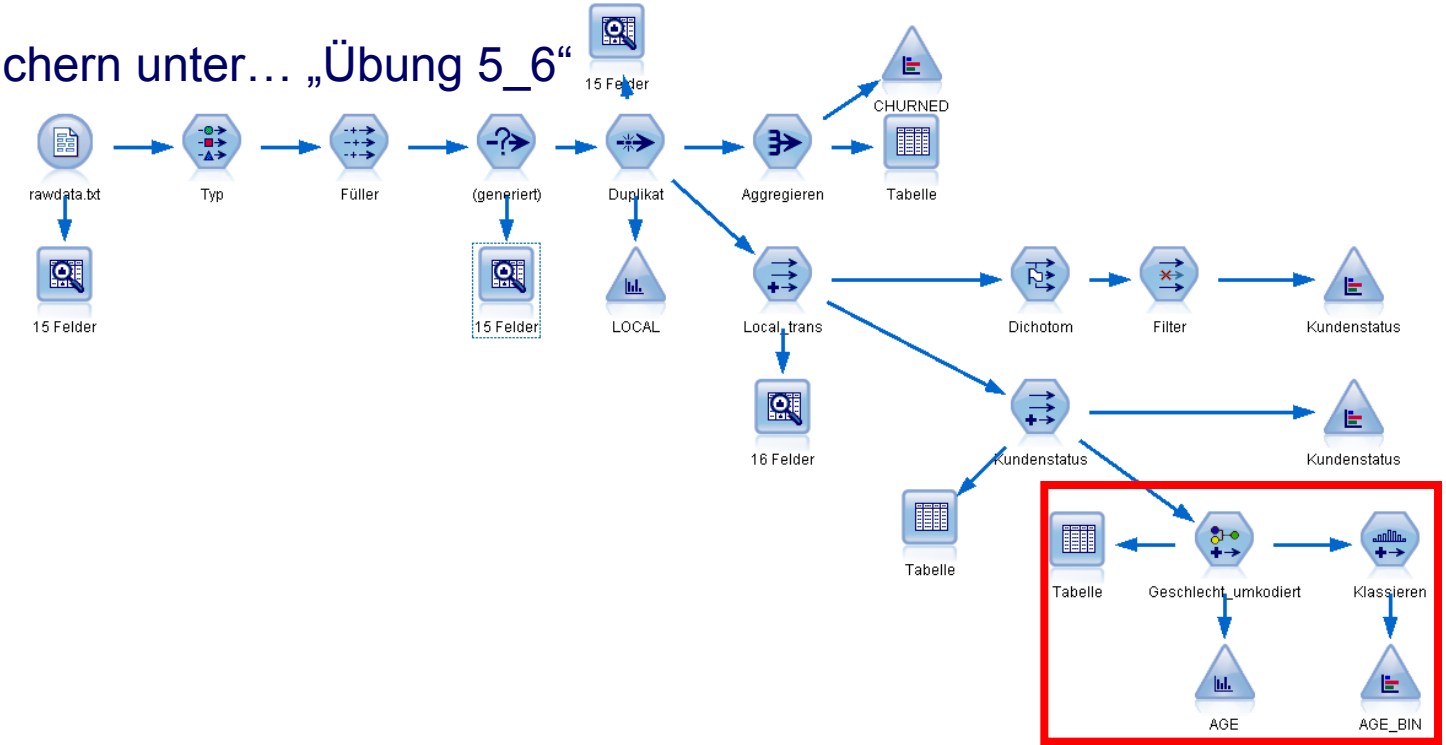
Wert	Anteil	%	Anzahl
0		56,33	832
1		43,67	645

OK

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.6 (1/2)

- Führen Sie eine Umkodierung des Feldes „SEX“ durch. Der Wert F soll mit 1, der Wert M mit 2 ersetzt werden. Das neue Feld soll Geschlecht_umkodiert benannt werden.
- Filtern Sie das Feld „SEX“ aus dem Stream heraus.
- Führen Sie eine Klassierung des Feldes „AGE“ durch. Die Klassenbreite soll 10 betragen.
- Stream speichern unter… „Übung 5_6“



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.6 (2/2)

■ Outputs

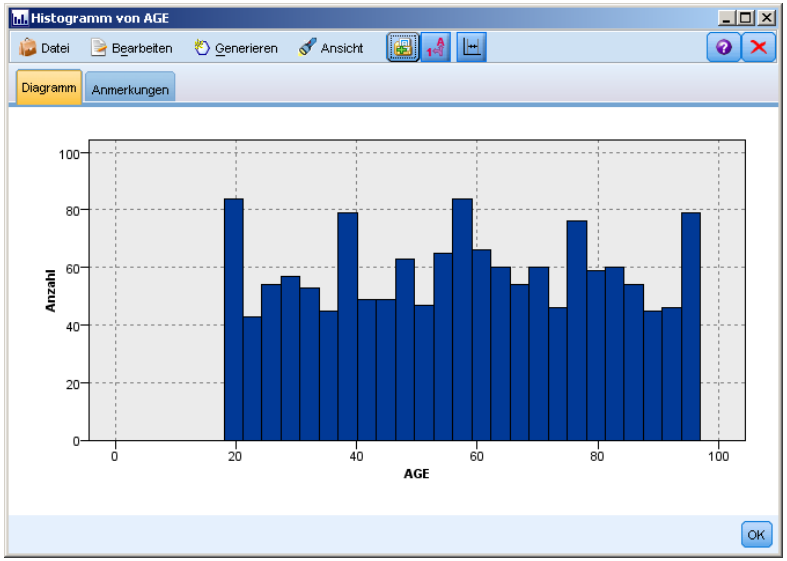
Tabelle (19 Felder, 1.477 Datensätze) #1

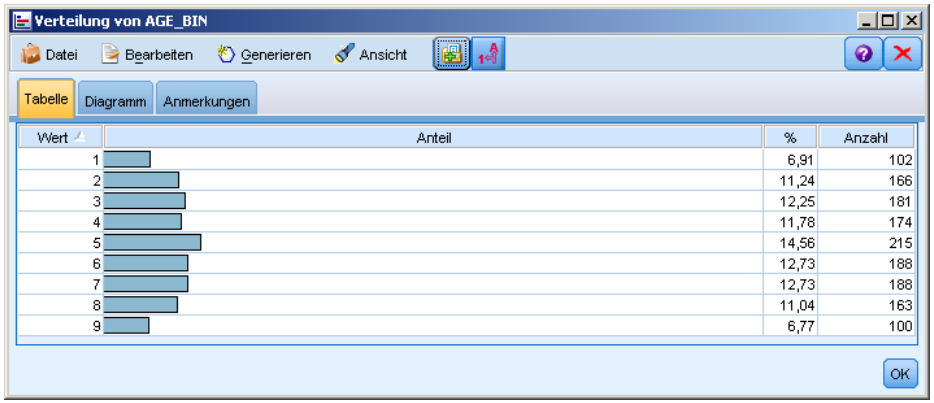
Datei Bearbeiten Generieren

Tabelle | Anmerkungen

	calBillType	LongDistanceBillType	AGE	SEX	Geschlecht_umkodiert	STATUS	CHILDREN	Est_Income
1	eeLocal	Standard	57	F	1	M	2	27535.300
2	eeLocal	Intnl_discount	60	M	2	M	2	83220.600
3	idget	Intnl_discount	52	M	2	S	0	88413.400
4	idget	Standard	53	M	2	M	1	34317.600
5	idget	Intnl_discount	86	F	1	M	1	99248.100
6	eeLocal	Intnl_discount	62	F	1	M	2	8749.800
7	idget	Intnl_discount	48	M	2	S	0	81420.000
8	idget	Intnl_discount	38	F	1	M	0	17499.200
9	idget	Standard	94	M	2	M	1	76213.200
10	idget	Intnl_discount	75	F	1	M	2	11434.600
11	eeLocal	Standard	56	F	1	S	2	27362.600
12	idget	Standard	77	F	1	M	2	37155.000
13	idget	Intnl_discount	97	M	2	M	2	75004.500
14	idget	Standard	32	M	2	M	0	54430.200
15	eeLocal	Intnl_discount	27	F	1	S	1	54370.700
16	eeLocal	Standard	90	M	2	S	2	83653.300
17	idget	Standard	59	F	M	S	0	23892.200
18	idget	Standard	77	F	1	S	0	34160.500
19	idget	Standard	27	M	2	M	2	50253.704
20	eeLocal	Intnl_discount	86	F	1	S	1	31639.200

OK



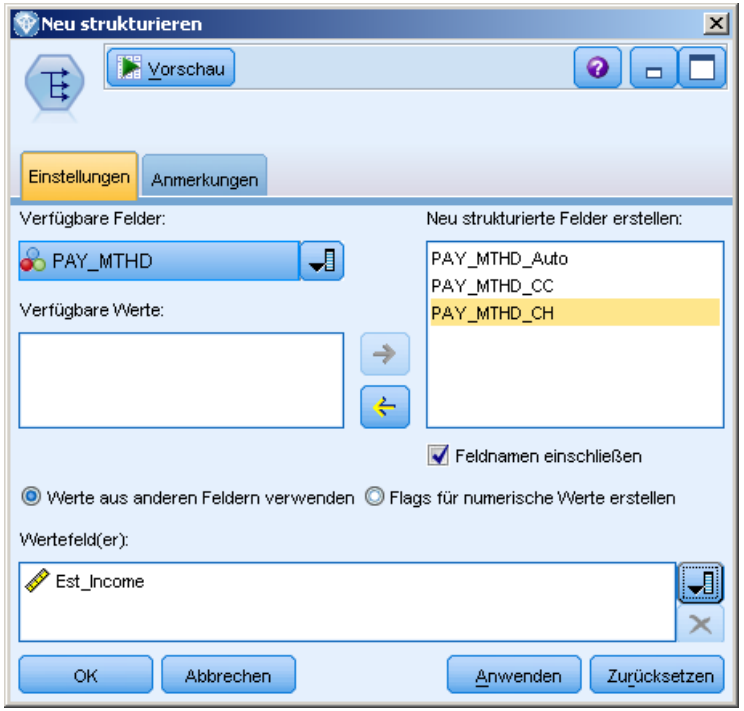


Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Umstrukturieren

Umsatzdaten werden oft in einem Datensatz pro Transaktion gespeichert, werden zur Analyse aber als Variablen nebeneinander (z. B. Umsatz pro Produktgruppe) benötigt. Dies lässt sich entweder mit einer Reihe von **Ableiten-Knoten** darstellen oder einfacher mit dem **Umstrukturieren-Knoten**.

- Nach entsprechender Umstrukturierung kann eine Aggregation zur Daten-reduktion durchgeführt werden.

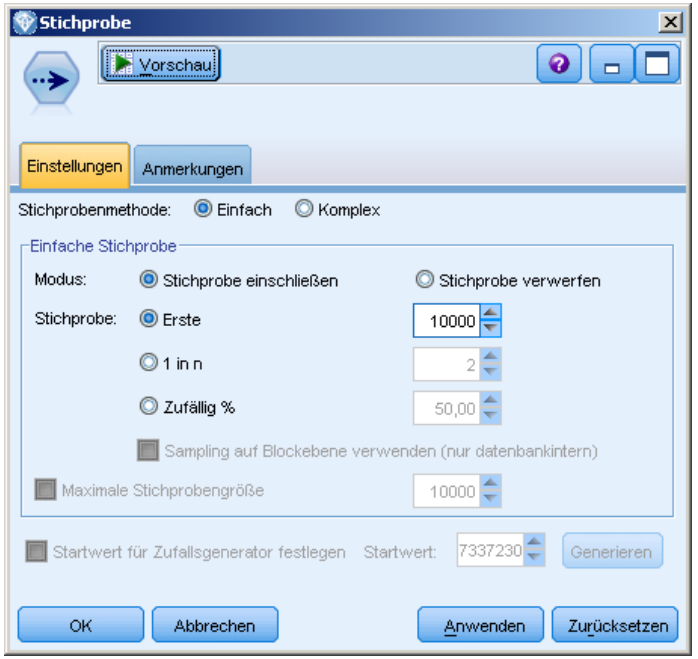


Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Stichproben

Stichprobe-Knoten

- Mit einem **Stichprobe-Konten** lassen sich Stichproben der Datensätze nach verschiedenen Verfahren bilden.
- Stichproben werden gebildet
  - zur Reduktion der Anzahl der Datensätze → Beschleunigung
  - zum Training und Test der Modelle (Stichwort Overfitting)
  - zum Vergleich von Modellen



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Partitionen zur Bildung von Data Mining-Modellen

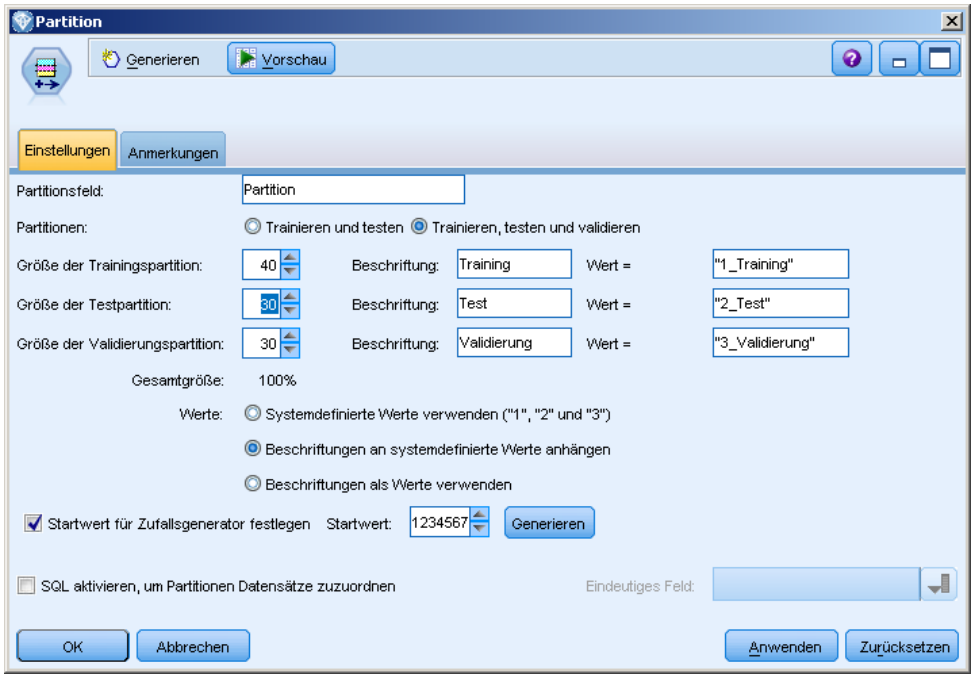
- Das Modell wird mit einem Teil der Daten trainiert (**Trainingsdaten**). Für das interessierende Merkmal sind die Ausprägungen bekannt und hiermit werden die Modellparameter bestimmt. Muster in den Daten werden erkannt.
- Mit einem weiteren Auszug aus der Datenbasis (**Testdaten**) wird das Modell durch die Bestimmung von Methodenparametern verfeinert bzw. die optimale Komplexität ermittelt. Dies ist notwendig, um zu verhindern, dass das Modell zu sehr die Spezifika der Trainingsdaten abbildet (Overfitting). Das Modell erhält einen allgemeineren Charakter und findet mit neuen unbekannten Daten bessere Lösungen (Generalisierungsfähigkeit).
- Die Leistungsfähigkeit des Modells bzw. aller angewandten Modelle kann mit einem dritten Teil der Ausgangsdaten abgeschätzt werden (sog. **Validierungsdaten** bzw. **Evaluationsdaten**). Dieser Teil der Daten ist klar von den ersten beiden zu trennen und soll nicht mit in die Modellbildung einfließen.
- Das Modell wird auf die **Auswertungsdaten** (neue Daten) angewandt. Die Ausprägungen des interessierenden Merkmals sind im Vorfeld nicht bekannt, sondern werden durch das Modell geschätzt. Die Vorhersagewerte werden als Grundlage für Entscheidungen genutzt.

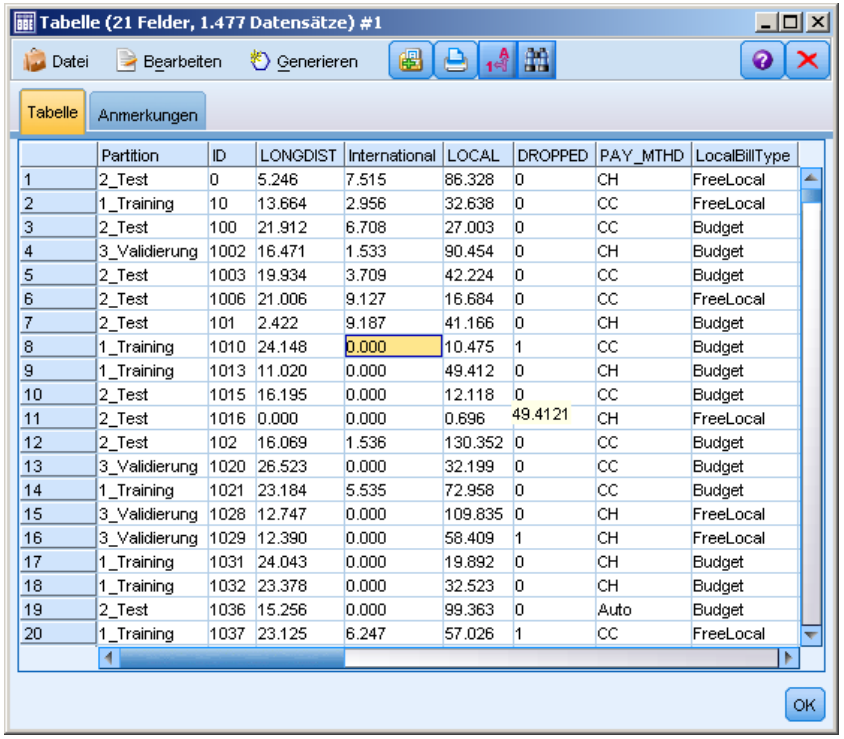
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Quelle: Berry/Linoff (2000): Mastering Data Mining, S. 185 ff.

Partitionen

Zufällige Partitionen für Training, Test und Validierung können mit einem **Partition-Knoten** gebildet werden.



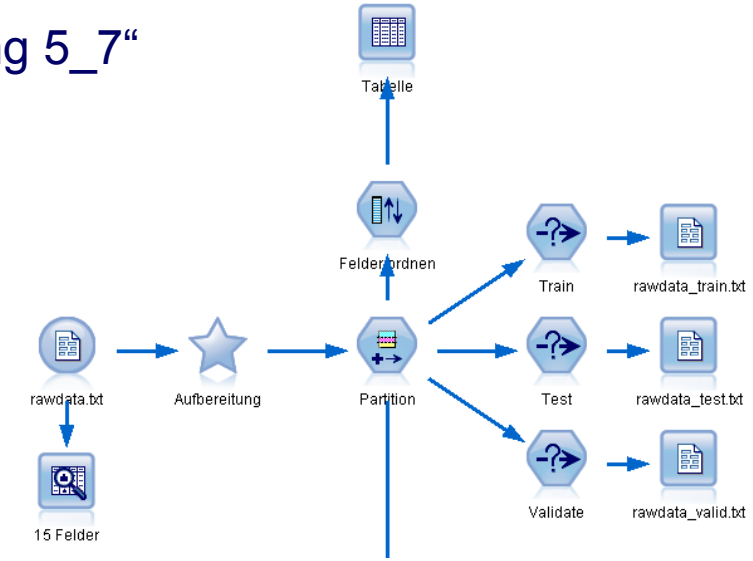


	Partition	ID	LONGDIST	International	LOCAL	DROPPED	PAY_MTHD	LocalBillType
1	2_Test	0	5.246	7.515	86.328	0	CH	FreeLocal
2	1_Training	10	13.664	2.956	32.638	0	CC	FreeLocal
3	2_Test	100	21.912	6.708	27.003	0	CC	Budget
4	3_Validierung	1002	16.471	1.533	90.454	0	CH	Budget
5	2_Test	1003	19.934	3.709	42.224	0	CC	Budget
6	2_Test	1006	21.006	9.127	16.684	0	CC	FreeLocal
7	2_Test	101	2.422	9.187	41.166	0	CH	Budget
8	1_Training	1010	24.148	0.000	10.475	1	CC	Budget
9	1_Training	1013	11.020	0.000	49.412	0	CH	Budget
10	2_Test	1015	16.195	0.000	12.118	0	CC	Budget
11	2_Test	1016	0.000	0.000	0.696	49.4121	CH	FreeLocal
12	2_Test	102	16.069	1.536	130.352	0	CC	Budget
13	3_Validierung	1020	26.523	0.000	32.199	0	CC	Budget
14	1_Training	1021	23.184	5.535	72.958	0	CC	Budget
15	3_Validierung	1028	12.747	0.000	109.835	0	CH	FreeLocal
16	3_Validierung	1029	12.390	0.000	58.409	1	CH	FreeLocal
17	1_Training	1031	24.043	0.000	19.892	0	CH	Budget
18	1_Training	1032	23.378	0.000	32.523	0	CH	Budget
19	2_Test	1036	15.256	0.000	99.363	0	Auto	Budget
20	1_Training	1037	23.125	6.247	57.026	1	CC	FreeLocal

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.7

- Ergänzen Sie den Stream um einen Partitionsknoten.
- Untergliedern Sie die Daten in eine Trainings- (40 %), Test- (30 %) und Validierungsmenge (30 %).
- Fassen Sie alle Knoten zwischen dem Quellknoten und dem Partitionsknoten in einen Superknoten zusammen.
- Exportieren Sie die Datensätze der Trainings-, Test- und Validierungsmenge in die entsprechenden Dateien rawdata_train.txt, rawdata_test.txt und rawdata_valid.txt (Feldtrennzeichen: „**;**“).
- Stream speichern unter … „Übung 5_7“



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenanreicherung

Anhängen-Knoten

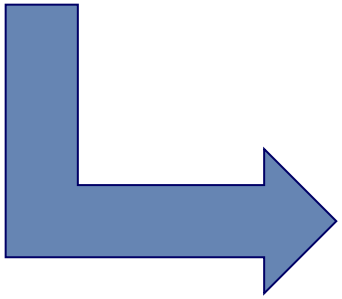
- Der **Anhängen-Knoten** verbindet mehrere Datenquellen miteinander. Dabei werden die einzelnen Datensätze nacheinander eingelesen und untereinander geschrieben.
- Anhängen-Knoten sind sinnvoll für die Kombination von Daten-Sets mit ähnlicher Struktur, aber unterschiedlichen Daten. Sie könnten beispielsweise Transaktionsdaten für verschiedene Zeiträume in verschiedenen Dateien gespeichert haben (z. B. zwei Absatzdatendateien für März und April). Angenommen, diese Dateien weisen dieselbe Struktur (dieselben Felder in derselben Reihenfolge) auf, werden sie mit dem Anhänge-Knoten in einer großen Datei zusammengefasst, die anschließend analysiert werden kann.
- Hat eine Datei mehr Felder als die erste Datei, werden diese automatisch mit field7, field8 ... benannt. Sind weniger Felder vorhanden, werden diese mit $null$ aufgefüllt.
- Bei der Zusammenführung der Daten müssen die einzelnen Felder das gleiche Skalenniveau aufweisen. Diese Einstellungen werden im Typ-Knoten vorgenommen (s. o.).

Datenanreicherung

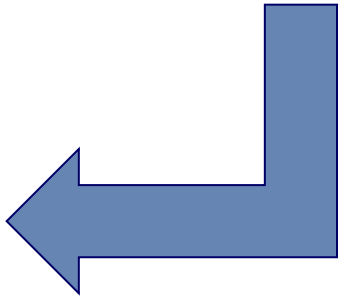
Anhängen-Knoten

ID	Tel-Vertrag	Hardware
134652346	50	Galaxy S4
234462645	200	iPhone 5
356835688	120	Xperia Z
467945679	50	HTC One

ID	Tel-Vertrag	Hardware
235564575	120	LG Optimus G
134627627	200	HTC One X+
264796745	50	iPhone 4s
226373567	50	Galaxy Note 2



ID	Tel-Vertrag	Hardware
134652346	50	Galaxy S4
234462645	200	iPhone 5
356835688	120	Xperia Z
467945679	50	HTC One
235564575	120	LG Optimus G
134627627	200	HTC One X+
264796745	50	iPhone 4s
226373567	50	Galaxy Note 2



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenanreicherung

Zusammenführen-Knoten

- Mit dem **Zusammenführen-Knoten** können weitere Datenfelder zu einem Datensatz hinzugefügt werden.

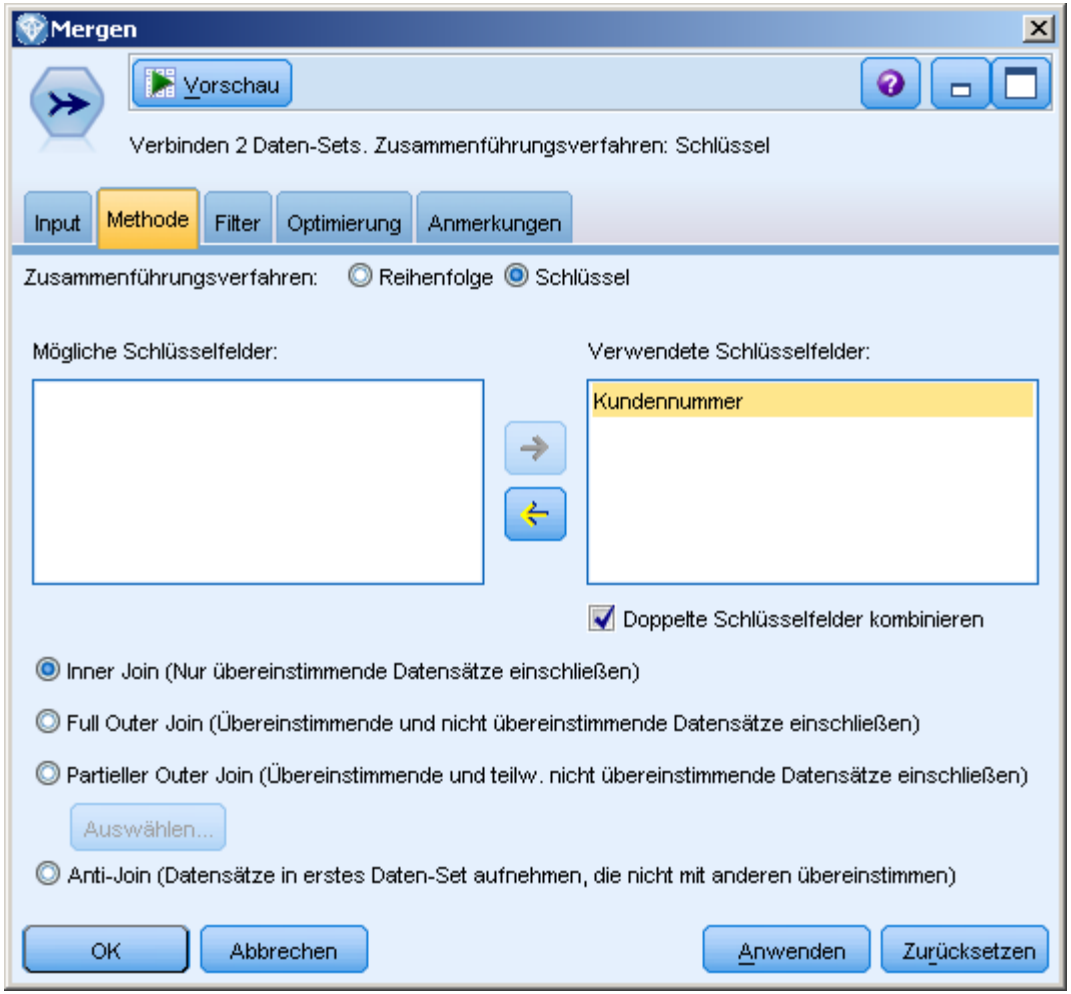
ID	Tel-Vertrag	Hardware
134652346	50	Galaxy S4
234462645	200	iPhone 5
356835688	120	Xperia Z
467945679	50	HTC One

ID	Dat-Vertrag	Usage
134652346	S	low
234462645	XL	high
893835745	S	normal
467945679	M	high

ID	Tel-Vertrag	Hardware	Dat-Vertrag	Usage
134652346	50	Galaxy S4	S	low
234462645	200	iPhone 5	XL	high
356835688	120	Xperia Z	$null$	$null$
467945679	50	HTC One	M	High
893835745	$null$	$null$	S	normal

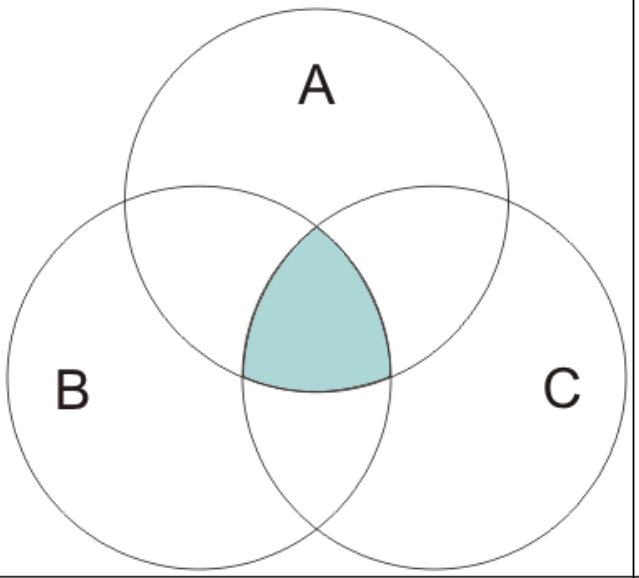
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenanreicherung

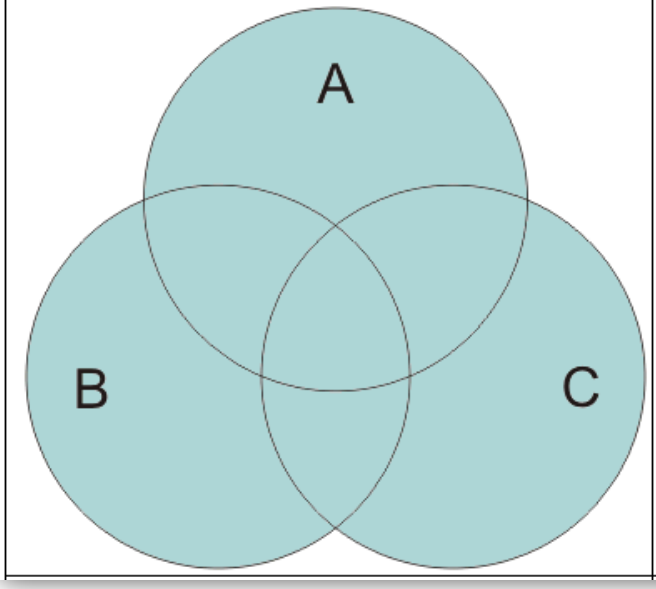


Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenanreicherung



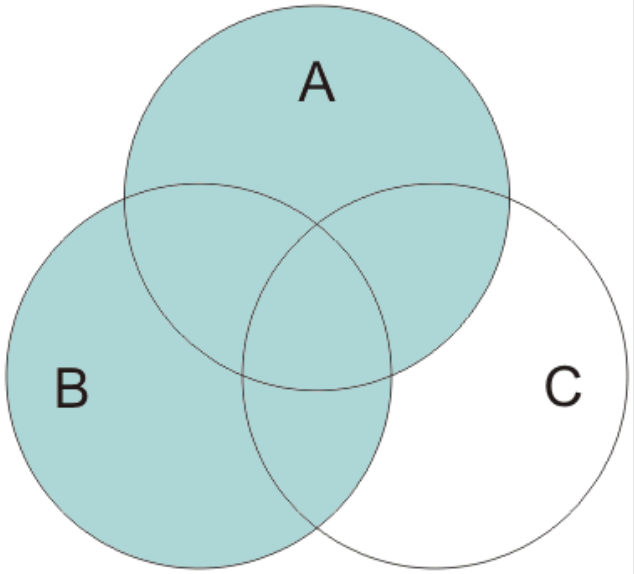
Ein **Inner Join** enthält nur Datensätze, bei denen ein Wert für das Schlüsselfeld bei allen Eingabetabellen gleich ist. Nicht übereinstimmende Datensätze werden nicht in das Ausgabe-Daten-Set aufgenommen.



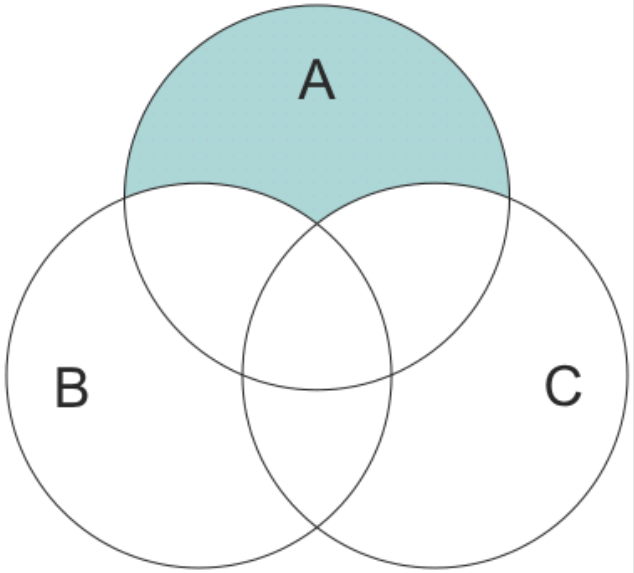
Bei einem **Full Outer Join** werden alle Datensätze (übereinstimmend und nicht übereinstimmend) aus den Eingabetabellen eingeschlossen. Linke und rechte Outer Joins werden als partielle Outer Joins bezeichnet und werden im Folgenden beschrieben.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Datenanreicherung



Ein **partieller Outer Join** enthält alle Datensätze, deren Übereinstimmung anhand des Schlüsselfeldes abgeglichen wurde, sowie nicht übereinstimmende Datensätze aus den angegebenen Tabellen. (Oder anders gesagt: Alle Datensätze aus bestimmten Tabellen und nur passende Datensätze aus anderen Tabellen.) Tabellen (wie beispielsweise A und B in der Abbildung) können mithilfe der Schaltfläche "Auswahl" auf der Registerkarte "Verbinden" ausgewählt werden. Partielle Joins werden auch linke bzw. rechte Outer Joins genannt, wenn nur zwei Tabellen zusammengeführt werden. Da Clementine die Zusammenführung von mehr als zwei Tabellen erlaubt, wird dieser Vorgang hier als partieller Outer Join bezeichnet.

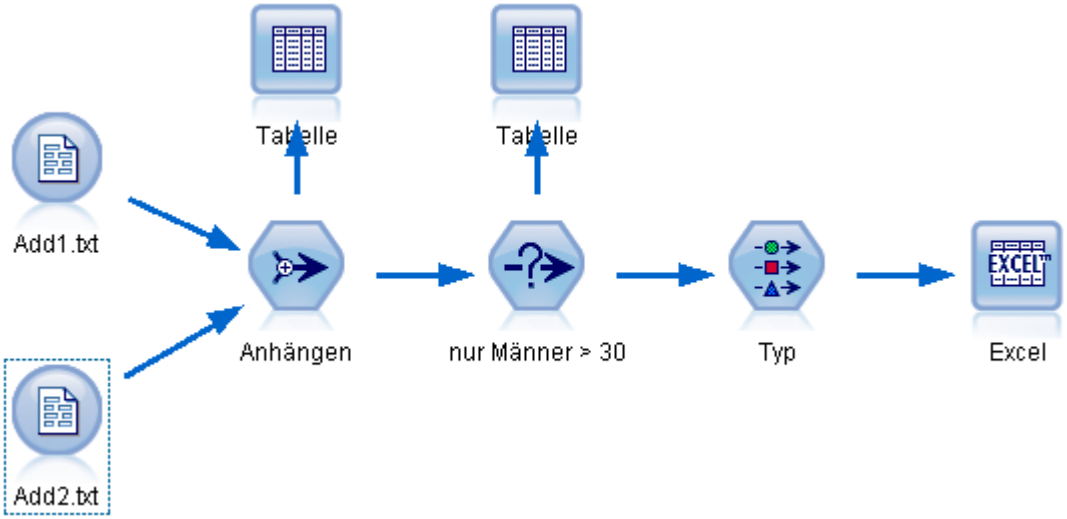


Bei **Anti-Join** werden nur nicht übereinstimmende Datensätze für die erste Eingabetabelle (Tabelle A in der Abbildung) aufgenommen. Bei diesem Join-Typ handelt es sich um das Gegenteil eines Inner Join. Es werden keine vollständigen Datensätze in das Ausgabe-Daten-Set aufgenommen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.8

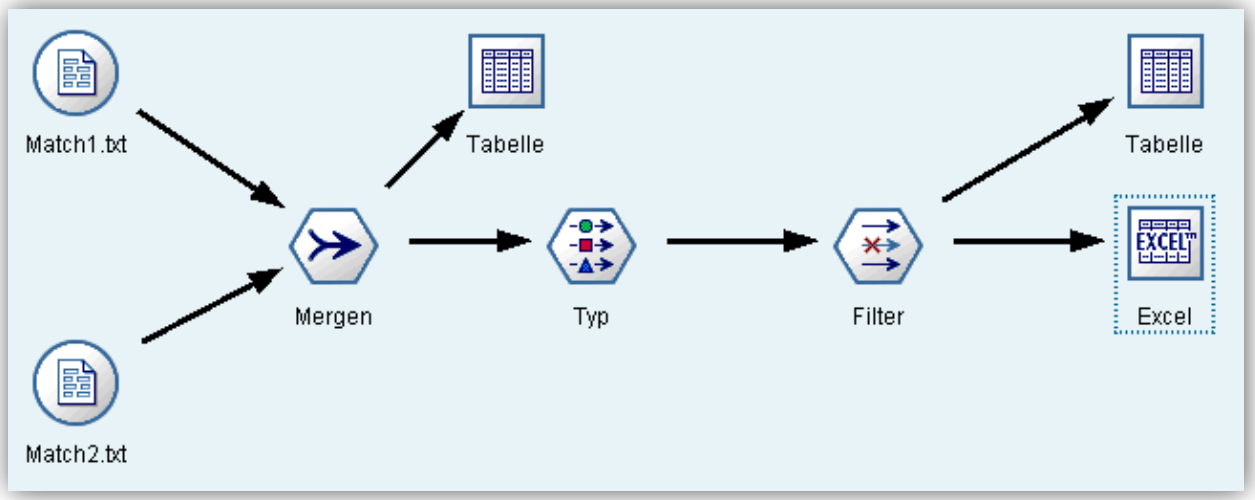
- Die Dateien Add1.txt und Add2.txt enthalten Personaldaten einer Firma.
- Führen Sie die beiden Dateien Add1.txt und Add2.txt über einen Anhängeknoten zusammen. Funktioniert das Anhängen über Namen und Position der Datenfelder?
- Lassen Sie sich in einer Tabelle die Daten des männlichen Personals über 30 Jahre ausgeben.
- Speichern Sie die Tabelle in einem Excel-File „Personal_angehängt.xlsx“.
- Stream speichern unter… „Übung 5_8“



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 5.9

- Die Dateien Match1.txt und Match2.txt enthalten Daten zu Filialen und Kundeneinkäufen.
- Verbinden Sie die beiden Dateien Match1.txt und Match2.txt über einen Zusammenführen-Knoten.
  Im Stream sollen dabei nur vollständige Datensätze verbleiben.
- Führen Sie eine Umbenennung des Felders „erzielter Umsatz“ in „Umsatz“ durch.
- Kontrollieren Sie die Daten über einen Tabellen-Knoten und speichern Sie die Daten in einem Excel-File „Kundenumsätze_verbunden.xlsx“.
- Stream speichern unter… „Übung 5_9“



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

1. Analytisches CRM
2. IBM SPSS Modeler
3. Knoten-Übersicht
4. Einlesen von Daten
5. Datenaufbereitung
6. Data Mining-Modelle

Test von Data Mining-Modellen

- **Methodenspezifische Tests:**
  Dienen primär der Optimierung der Methodenparameter und der Konzeption leistungsfähiger Data Mining-Modelle auf der Grundlage **einer** spezifischen Data Mining-Methode.
- **Methodenunabhängige Tests:**
  - Sind auf alle Data Mining-Methoden anwendbar
    (z.B. Test auf Approximations- und Prognosetauglichkeit bzw. auf Generalisierungs-fähigkeit)
  - Dienen primär dem Vergleich **unterschiedlicher** Data Mining-Methoden im Rahmen der Auswahl oder Kombination von Data Mining-Methoden.

Quelle: Hippner et. al. 2001, S. 72 ff.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Evaluation - das richtige Data Mining-Modell auswählen

- **Problem:** Je nach Aufgabenstellung kommen unterschiedliche Data Mining-Methoden zur Erstellung einer Prognose in Frage. Welche passt am besten?
- **Lösung:**
  1. Für jede Data Mining-Methode mithilfe der selben Trainings- und Testdaten ein oder mehrere Modelle erstellen.
  2. Die verschiedenen Modelle werden jeweils mit den Evaluationsdaten geprüft. Für Regressionsprobleme eignen sich Fehlermaße wie der mittlere absolute Fehler. Für Klassifikationsprobleme lässt sich für jedes Modell eine **Trennschärfetabelle** anfertigen, die auf einen Blick jeweils die Zahl der Fehlklassifikationen anzeigt.
- **Beispiel:** Prognose j/n

		Evaluationsdaten	
		ja	nein
Vorhersage Modell 1	ja	2%	4%
	nein	12%	82%

		Evaluationsdaten	
		ja	nein
Vorhersage Modell 2	ja	8%	24%
	nein	6%	62%

Quelle: Berry/Linoff 2000, S. 54 ff.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Evaluation - Bewertung von Trennschärfetabellen

- In unserem Beispiel sind **zwei Arten von Fehlklassifikationen** zu unterscheiden: Entweder wird für Personen, bei denen die Merkmalsausprägung „ja“ zutrifft, fälschlicherweise „nein“ vorhergesagt oder umgekehrt.
- Meist lassen sich bei der Bewertung von Data Mining-Modellen die beiden Fehlklassifikationsarten gewichten - bspw. nach den unterschiedlichen Kosten einer Fehlklassifikation.
- **Beispiel:**
  - Bei medizinischen Krebsuntersuchungen muss man sich auf die Aussage „nein“ eines Krebserkennungstests verlassen können. Lieber würde man im Zweifel weitere Tests veranlassen, als in Kauf zu nehmen, dass die Krankheit bei Patienten unentdeckt bleibt.
  - Anders verhält es sich bei Vorhersagen, welche Personen auf ein Werbe-Mailing reagieren werden. Hier sind für Personen, die fälschlicherweise als Reagierer eingestuft werden jeweils die Kosten eines Mailings anzusetzen und für Reagierer, die gar nicht erst angeschrieben werden, die durchschnittlich entgangenen Gewinne.

Quelle: Berry/Linoff 2000, S. 54 ff.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Evaluation - Gains Chart (Cumulated Response-Chart)

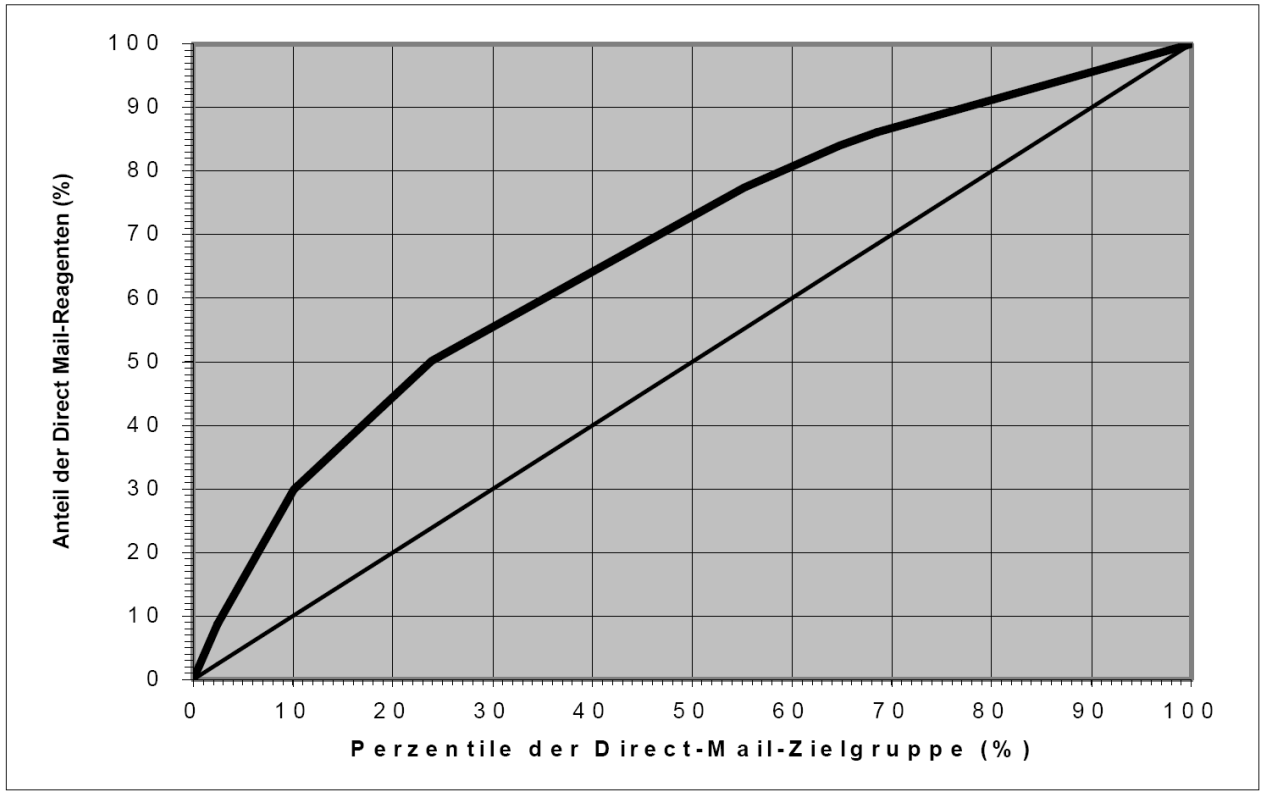
- Visualisiert den Nutzen von Data Mining-Modellen.
- Wie beim Lift Chart werden die Datensätze absteigend angeordnet, so dass die Elemente mit größerer Zugehörigkeitswahrscheinlichkeit zur interessierenden Klasse links stehen (Klassifikationsmodell).
- Die Kurve des Gains Chart zeigt an, wie hoch der Anteil der richtig klassifizierten Datensätze an der Gesamtzahl der tatsächlich zur interessierenden Klasse gehörigen Datensätze bei der Anwendung des Modells auf eine bestimmte Anzahl der Ausgangsdaten bereits ist.
- Beispielsweise könnte die Klasse der Reagierer eine interessierende Klasse darstellen.

Quelle: Hippner et. al. 2001, S. 79

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Gains Chart – Beispiel

- Aus dem Chart lässt sich ablesen, dass die vom Klassifikationsmodell hinsichtlich ihrer Reaktionswahrscheinlichkeit am besten bewerteten 25 % der Datensätze des Evaluierungsdatensatzes bereits 50 % aller tatsächlichen Reagierer abdecken.

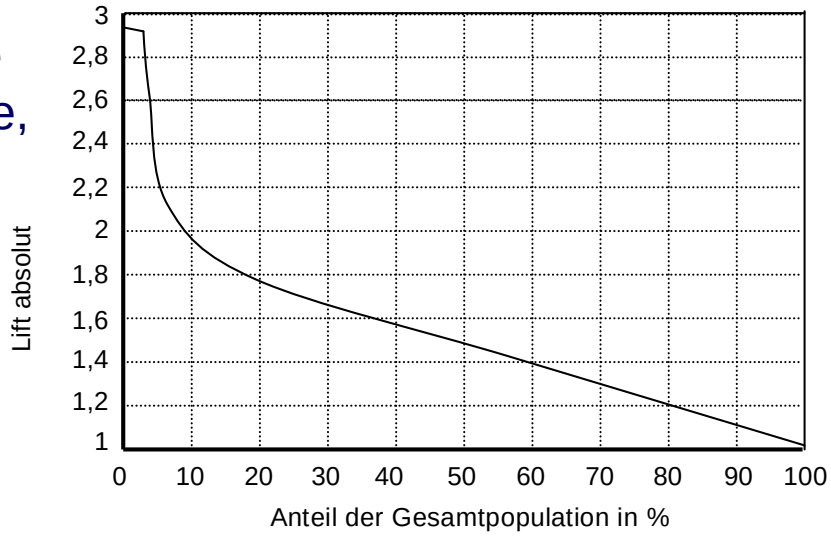


Quelle: Hippner et. al. 2001, S. 79

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Evaluation – Lift Chart

- Um unterschiedliche Data Mining-Modelle beurteilen und eines davon für die Auswertungsdaten auswählen zu können, lässt sich auch ein Lift Chart anfertigen.
- Ein **Lift Chart** beschreibt, wie stark die Konzentration der interessierenden Klasse durch das Modell gesteigert werden konnte, im Vergleich zu einer zufälligen Auswahl eines bestimmten Anteils der Datensätze (Klassifikation). Auf der horizontalen Achse wird der Anteil aller Datensätze dargestellt (in %). Zugleich repräsentiert diese Achse einen konstanten Lift-Wert von 1, der bei zufälliger Auswahl aus der Gesamtpopulation zu erwarten wäre.



- Gains- und Lift-Charts können alle bestehenden Modelle berücksichtigen. Diese werden durch separate Kurven dargestellt. Anhand der Kurvenführung lässt sich das am besten geeignete Modell auswählen.

Quelle: Berry/Linoff (2000), S. 54 ff.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&R

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

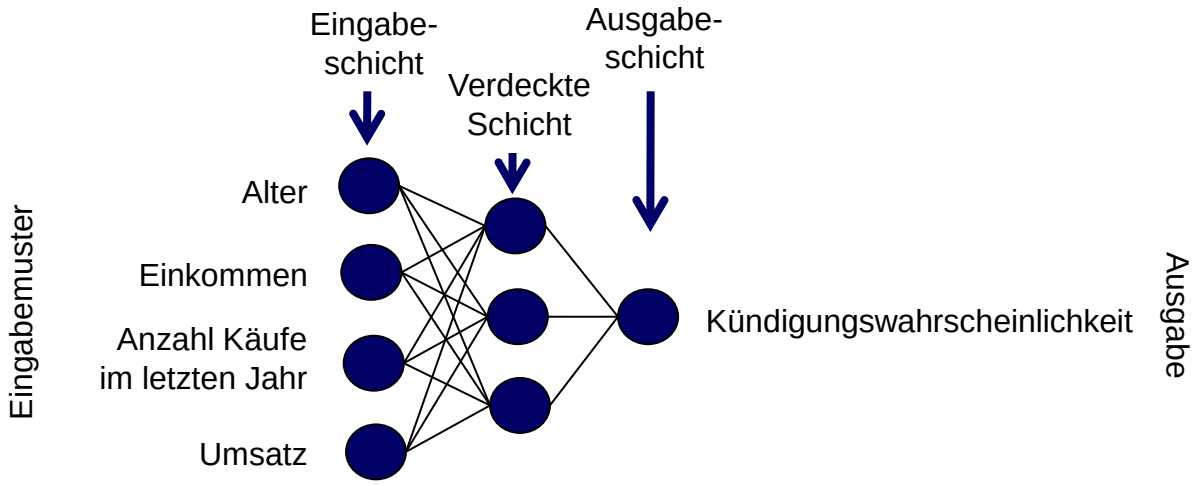
6.4.2 Two-Step

6.5 Assoziationsanalyse

6.5.1 A Priori

Grundlagen

Der Netzwerk-Knoten wird verwendet um Multi-Layer-Perzeptron (MLP)-Netze zu erzeugen. **MLP-Netze** bestehen aus Schichten von Neuronen, bei denen jedes Neuron mit allen Neuronen der vorherigen Schicht mit gewichteten Verbindungen in Kontakt steht. Sie bestehen aus einer Eingabeschicht, einer Ausgabeschicht und einer oder mehreren versteckten Schichten.



- **Radiale Basisfunktionsnetze (RBFN)** werden im Rahmen dieser Übung nicht behandelt.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Grundlagen

- **Voraussetzungen im Modeler:**
  - Der Netzwerk-Knoten erwartet mindestens ein Feld mit der Richtung Prädiktor (Input) und mindestens ein Feld mit der Richtung Ziel (Output).
  - Die hier verwendeten Netze haben keine Restriktionen bzgl. der Feld-Typen. Es können alle Skalenniveaus verarbeitet werden (Achtung: Keine Modellrestriktionen bedeutet nicht, dass keine Aufbereitung – z. B. Umkodierung – notwendig ist!).
- Ursprünglich sind alle Gewichtungen zufällig und die Antworten, die vom Netz stammen, sind wahrscheinlich unsinnig. Das Netz lernt durch **Training**. Datensätze, deren Zielwerte bekannt sind, werden dem Netz wiederholt präsentiert und die Antworten werden mit den bekannten Zielwerten verglichen.
- Die Gewichtungen werden schrittweise geändert, sobald eine falsche Vorhersage erfolgt. Je weiter das Training fortschreitet, desto genauer wird das Netz bei der Replizierung der bekannten Ergebnisse. Das Netz verbessert seine Vorhersagen so lange, bis mindestens eines der Grenzkriterien erfüllt ist. Nachdem das Netz trainiert ist, kann es auf zukünftige Fälle angewendet werden, bei denen das Ergebnis unbekannt ist.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Trainingsmethoden

- Der Netzwerk-Knoten stellt u. a. diese **Trainingsmethoden** zur Verfügung:
  - **Standard:** Bei dem Verfahren wird ein einziges Modell erstellt, um das Ziel unter Verwendung der Prädiktoren vorherzusagen. In der Regel gilt, dass Standardmodelle einfacher interpretiert und schneller gescort werden können, als mithilfe von Boosting, Bagging oder durch Koppelung mit einem Server erzielte Modelle.
  - **Boosting:** Bei dem Verfahren wird unter Einsatz der Verbesserung ein Ensemble-Modell erstellt. Dabei wird eine Modellsequenz erzeugt, um genauere Vorhersagen zu erhalten. Das Verfahren ist sehr langsam, liefert jedoch zumeist gute Ergebnisse.
  - **Bagging:** Bei dem Verfahren wird unter Einsatz der Verstärkung (Bootstrap-Aggregation) ein Ensemble-Modell erstellt. Dabei werden mehrere Modelle erzeugt, um zuverlässigere Vorhersagen zu erhalten.
  - **Server:** Bei dieser Methode wird ein Ensemble-Modell durch Aufteilen des Daten-Sets in separate Datenblöcke erstellt. Diese Option ist empfehlenswert, wenn Ihr Daten-Set zu groß für die Erstellung eines der oben erwähnten Modelle oder die inkrementelle Modellerstellung ist. Unter Umständen kann das Modell mit dieser Option schneller als ein Standardmodell erstellt werden, das Scoren dauert jedoch evtl. länger als bei einem Standardmodell. Diese Option erfordert eine SPSS Modeler Serververbindung.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

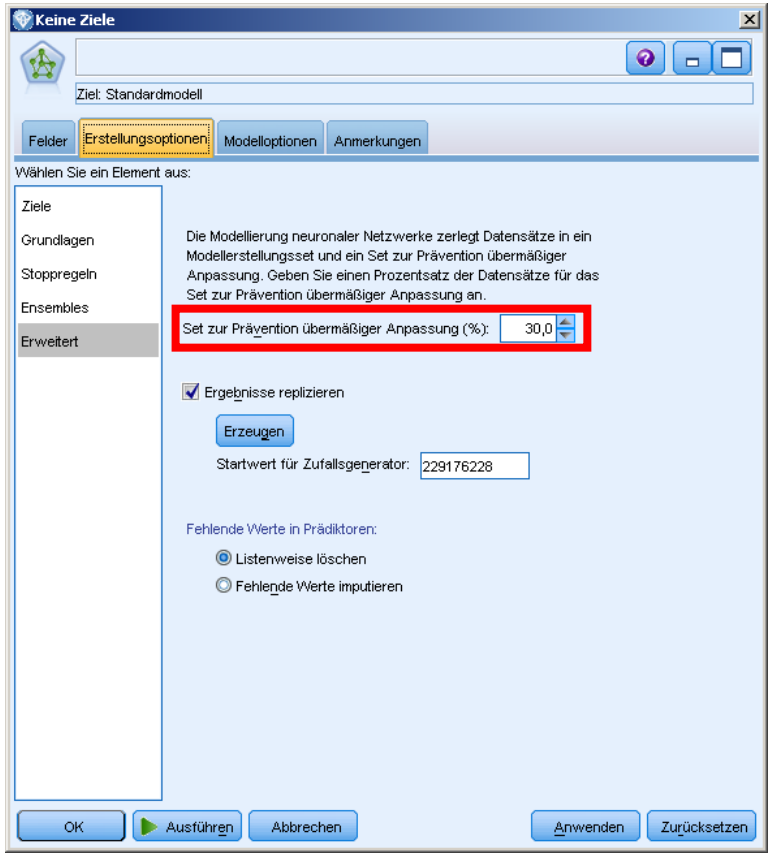
Trainingsmethoden

- **Wie findet man das beste Modell?**
  → Sowohl bei der Wahl der Trainingsmethode, als auch bei den jeweiligen Einstellungen gilt: mehrere Modelle ausprobieren!

- **Einige Anhaltspunkte zu den Trainingsmethoden:**
  - Wählen Sie die „Standard“-Methode, wenn nur wenig Zeit für die Analyse zur Verfügung steht.
  - Wenn die Genauigkeit höchste Priorität genießt und es keine Zeitrestriktion gibt, empfiehlt sich die „Boosting“- bzw. „Bagging“-Methode.
  - Der „Server“-Algorithmus wird selten verwendet. Dies kann schon aufgrund der Serveranforderung nur in großen Unternehmen erfolgen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übertrainieren verhindern

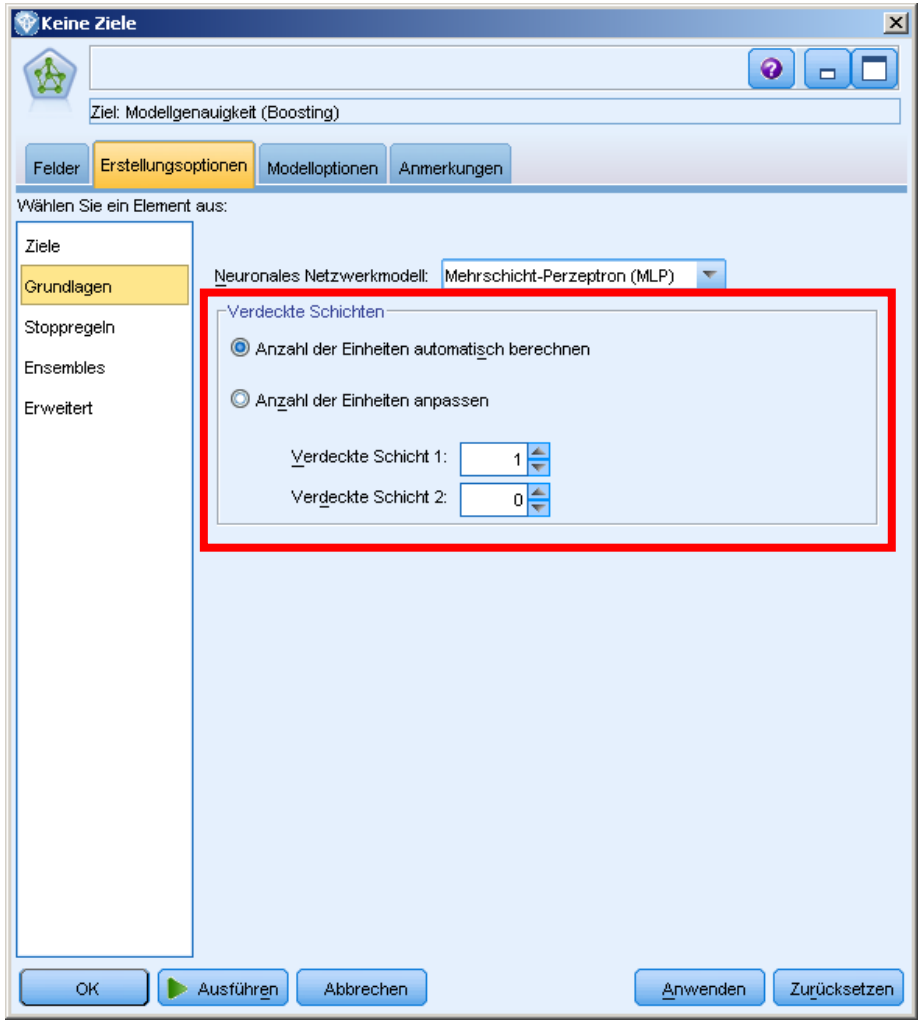


- Bei dieser Option werden die Daten zum Zwecke der Modellerstellung zufällig in **getrennte Trainings- und Test-Sets** aufgeteilt. Das Netzwerk wird mit dem Trainings-Set trainiert und die Genauigkeit wird mit dem Test-Set abgeschätzt. Den Anteil, der für den Test zu verwendenden Daten, geben Sie im Feld Set zur Prävention übermäßiger Anpassung (%) des Netzwerkknotens an. Der Rest der Daten wird für das Training verwendet.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Experten-Einstellungen

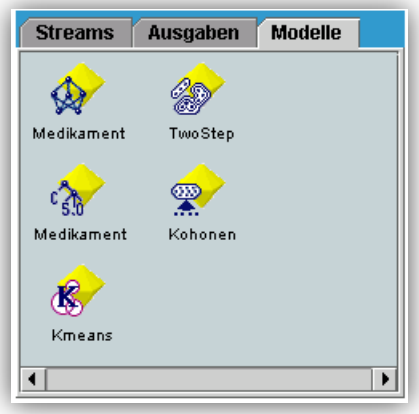
- Je nach verwendeter Trainingsmethode sind unterschiedliche **Experten-Einstellungen** möglich. Unter anderem kann die Anzahl der einzelnen Schichten vorgegeben werden.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Generierte Modelle

- Generierte Modelle sind das Ergebnis der Datenmodellierung
- Ein generierter Modellknoten wird bei jeder erfolgreichen Ausführung eines Modellierungsknotens erstellt.
- Die generierten Modelle werden in der **Palette der generierten Modelle** gespeichert (diese befindet sich auf der Registerkarte "Modelle" im Manager-Fenster rechts oben im Modeler-Fenster). Dort werden sie als diamantförmige Symbole dargestellt (und gelegentlich "Nuggets" genannt). Von dort können sie ausgewählt und durchsucht werden, um Details des Modells anzuzeigen.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Generierte Modelle

- Die generierten Modelle können in Streams eingebaut werden, um bspw. neue Daten mithilfe des Modells zu bewerten. Dazu:
  - In der Palette der generierten Modelle **doppelt** auf einen generierten Modellknoten klicken bzw. das Symbol bei gedrückter linker Maustaste in den Stream-Zeichenbereich „**hineinziehen**“.
  - Mit der rechten Maustaste auf einen generierten Modellknoten klicken und im Kontext-Menü „**Zu Stream hinzufügen**“ auswählen.

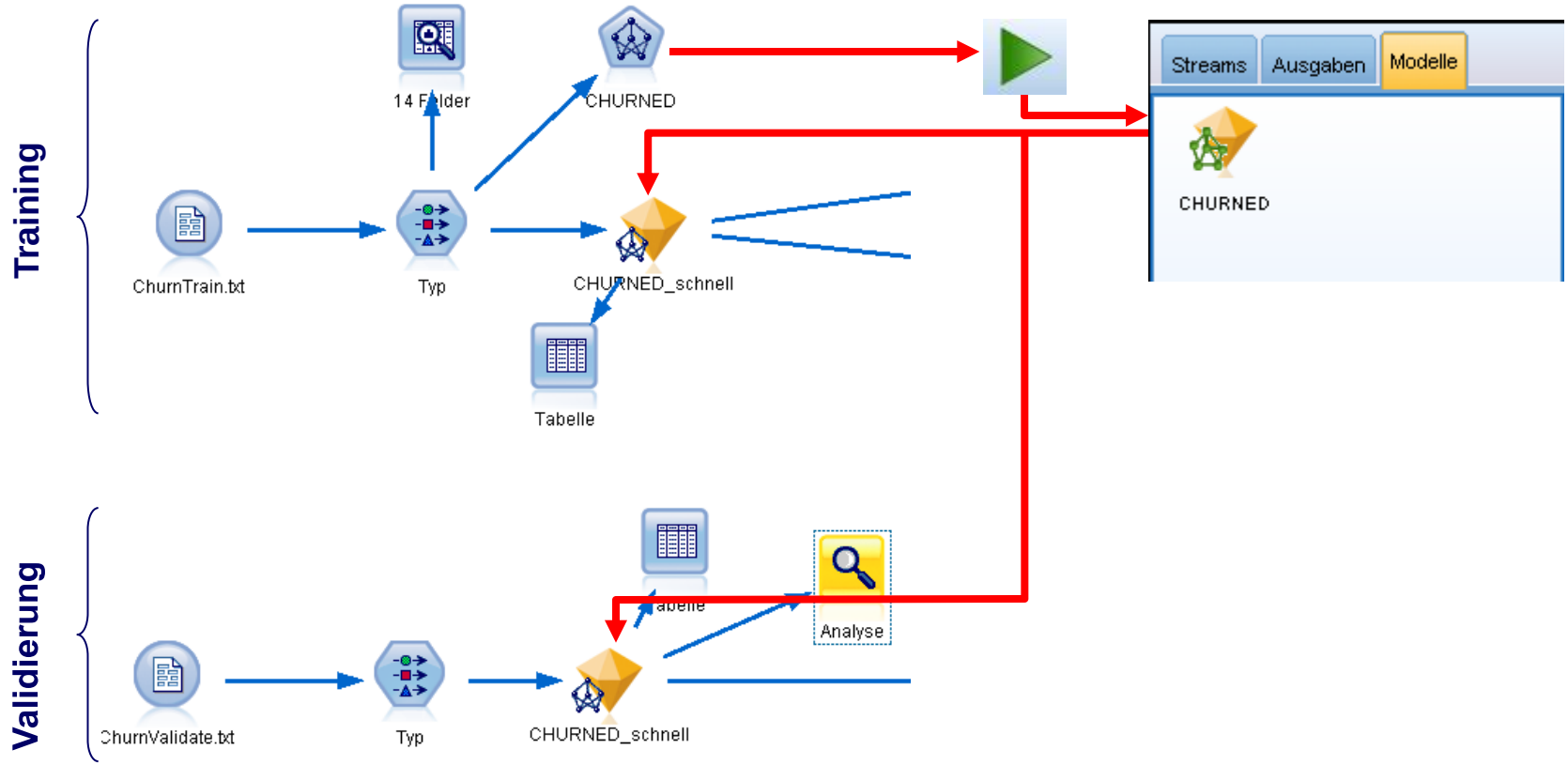
→ **Data Mining-Methode**

→ **Data Mining-Modell**

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Anwendungsbeispiel

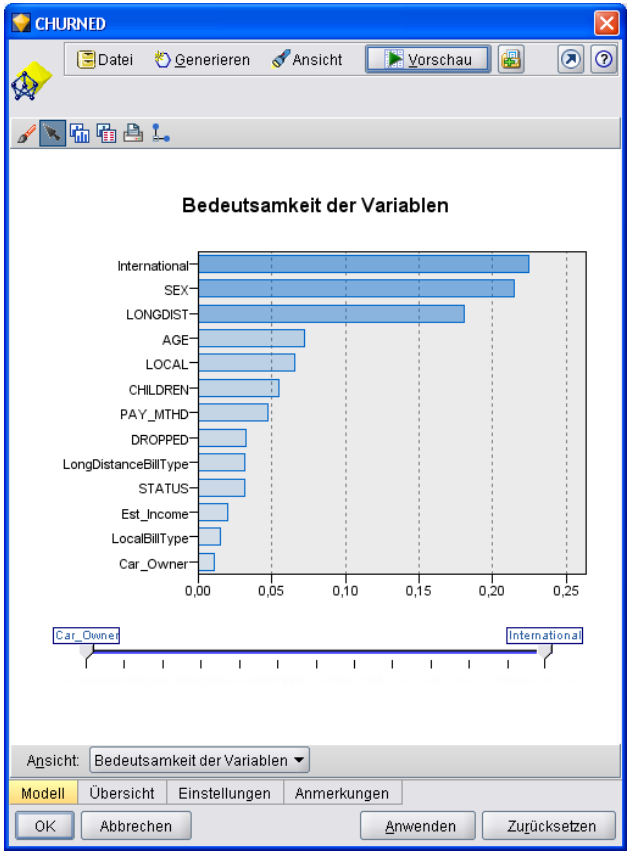
- Beim Testen des Modells wird der **gleiche generierte Modellknoten** verwendet, da genau dieses Modell untersucht („validiert“) werden soll.

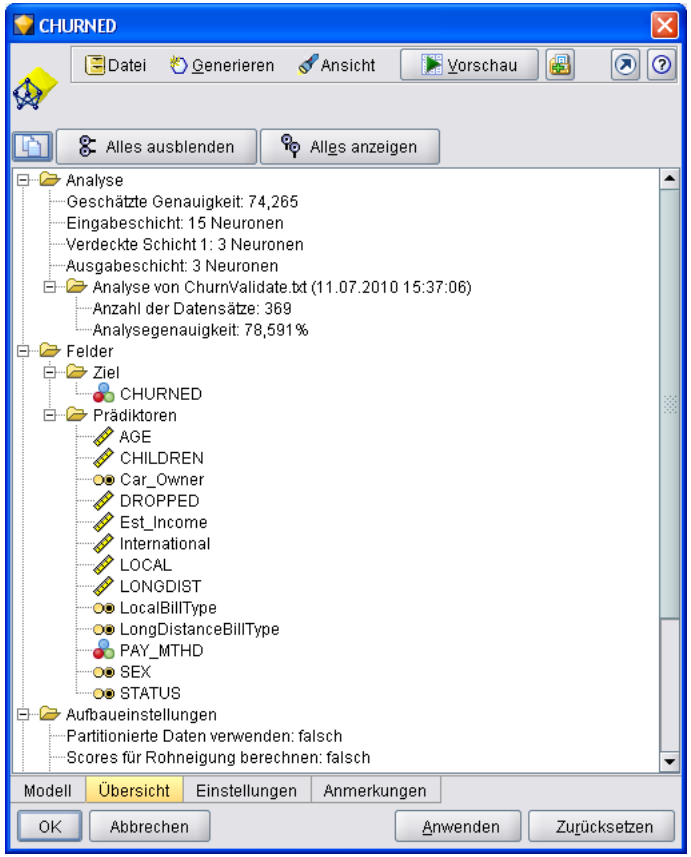


Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Anwendungsbeispiel

- Auf der Registerkarte „Übersicht“ eines generierten Modells werden Informationen über die Felder, die Aufbaueinstellungen und die Modellschätzung angezeigt. Die Ergebnisse werden in einer Baumansicht dargestellt, die durch Klicken auf bestimmte Elemente erweitert bzw. reduziert werden kann.





Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Anwendungsbeispiel

- In der Datenbasis werden zwei neue Datenfelder erzeugt:
  - **$N-Churned**
    Prognose des Netzwerks für das entsprechende Ausgabefeld
  - **$NC-Churned**
    Konfidenzwert der Prognose

Tabelle (17 Felder, 1.108 Datensätze)

Datei Bearbeiten Generieren

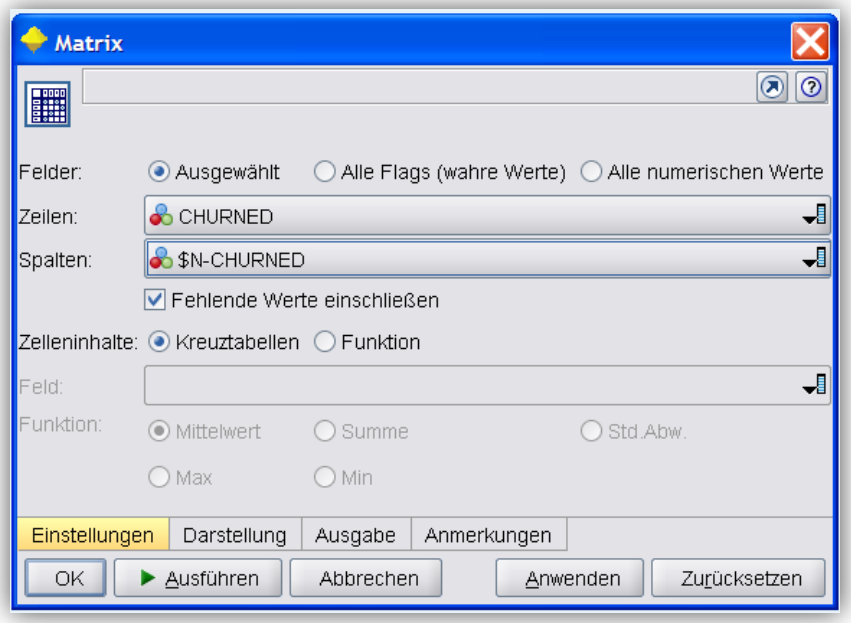
	ID	CHURNED	$N-CHURNED	$NC-CHURNED
1	0	Vol	Vol	0.669
2	3	InVol	InVol	0.931
3	4	Vol	Vol	0.572
4	10	Vol	Current	0.978
5	11	InVol	InVol	0.648
6	13	Vol	InVol	0.884
7	19	Vol	Vol	0.673
8	20	Vol	Vol	0.684
9	22	Current	Current	0.959
10	26	Vol	Current	0.753
11	28	Current	Vol	0.649
12	30	Current	Current	0.769
13	32	Vol	Vol	0.688
14	33	Current	Current	0.974
15	40	Current	Vol	0.643
16	49	Vol	Current	0.991
17	55	Current	Vol	0.599
18	59	Current	Current	0.990
19	65	Vol	Vol	0.645
20	69	InVol	InVol	0.281

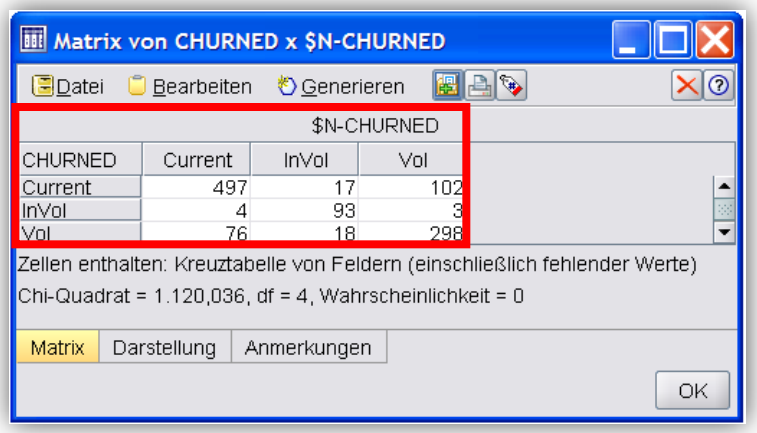
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Anwendungsbeispiel

- **Evaluierung des erzeugten Modells:**
  - Matrix-Knoten
  - Analyse-Knoten
  - Auswertung-Knoten

Matrix-Knoten: Vergleich der prognostizierten mit den tatsächlichen Werten. (Trennschärfetabelle)

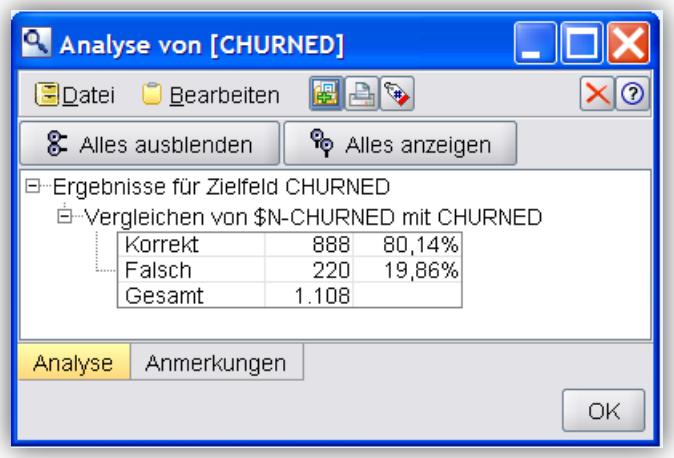



Matrix von CHURNED x $N-CHURNED
Datei Bearbeiten Generieren
$N-CHURNED
CHURNED Current InVol Vol
Current 497 17 102
InVol 4 93 3
Vol 76 18 298
Zellen enthalten: Kreuztabelle von Feldern (einschließlich fehlender Werte)
Chi-Quadrat = 1.120,036, df = 4, Wahrscheinlichkeit = 0
Matrix Darstellung Anmerkungen
OK

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Anwendungsbeispiel

Analyse-Knoten: Vergleicht die Gesamtzahl der tatsächlichen und der vorhergesagten Ausprägungen. Durch Aktivieren der Option auch die Ausgabe einer Trennschärfetabelle möglich.



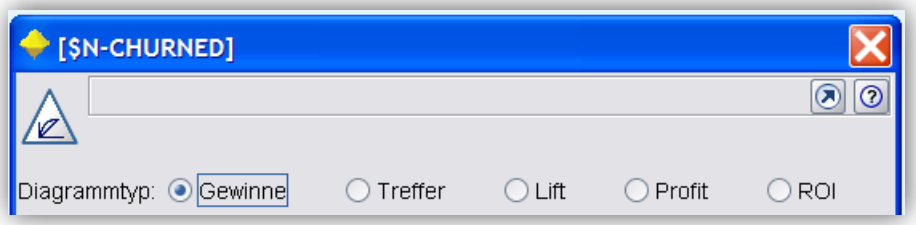
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Anwendungsbeispiel

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

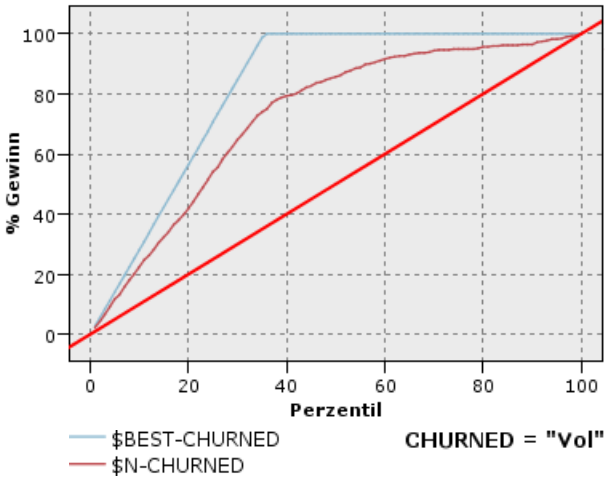
Auswertung-Knoten: Zeigt, wie gut Modelle bestimmte Ergebnisse vorhersagen. Die **Datensätze werden auf der Grundlage des vorhergesagten Wertes und des Konfidenzwertes für die Prognose sortiert**. Die Datensätze werden in gleich große Gruppen (Quantile) aufgeteilt. Anschließend wird der Wert des Geschäftskriteriums für jedes Quantil geplottet, vom höchsten Wert bis zum niedrigsten Wert. Mehrere Modelle werden als separate Linien im Plot dargestellt.

- **Evaluationsdiagrammtypen u. a.:**
  - **Gewinne**
    Gewinne sind definiert als der Anteil an allen Treffern, der in den einzelnen Quantilen vorliegt.
  - **Lift**
    Beim Lift wird der Prozentsatz der Datensätze in jedem Quantil, die als Treffer gelten, mit dem Gesamtprozentsatz der Treffer in den Trainingsdaten verglichen.

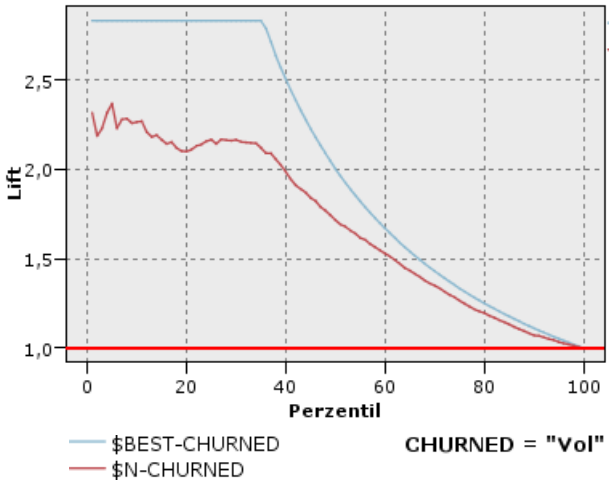


Anwendungsbeispiel

- **Gewinn-Diagramm (Gains Chart)**
  Kumulative Gewinndiagramme beginnen stets bei 0%, verlaufen von links nach rechts und enden bei 100%. In einem guten Modell steigt das Gewinndiagramm steil auf 100% an und verbleibt dann auf diesem Niveau. Bei einem Modell ohne Informationsgehalt verläuft eine diagonale Linie von links unten nach rechts oben.



- **Lift Chart**
  Kumulative Lift Charts beginnen in der Regel bei einem Wert über 1,0 und fallen von links nach rechts allmählich ab. Die rechte Kante des Diagramms entspricht dem gesamten Daten-Set; das Verhältnis der Treffer in den kumulativen Quantilen zu den Treffern in den Daten beträgt 1,0. Bei einem guten Modell sollte der Lift auf der linken Seite deutlich über 1,0 beginnen, von links nach rechts auf einem hohen Niveau verbleiben und dann auf der rechten Seite des Diagramms abrupt auf 1,0 fallen. Bei einem Modell ohne Informationsgehalt liegt die Linie im gesamten Diagramm bei einem Wert um 1,0.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.1 (1/2)

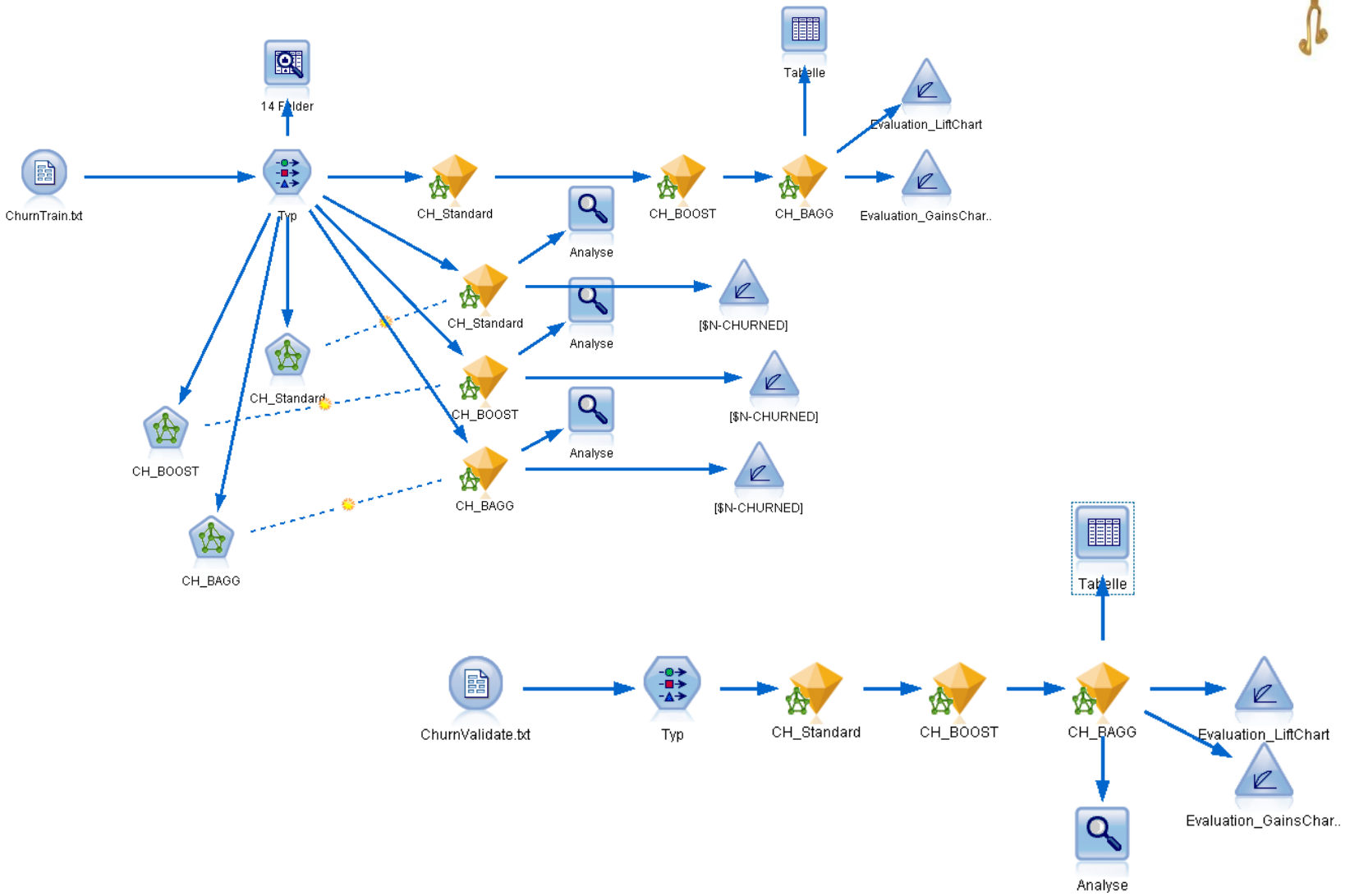
Abwanderungsgefährdete Kunden prognostizieren
(Dateien Churntrain.txt und Churnvalidate.txt)

- Ein Telekommunikations-Anbieter beauftragt Sie ein Modell zu entwickeln, das abwanderungsgefährdete Kunden prognostiziert.
- Legen Sie für jede Datei die Prädiktoren-Felder und das Zielfeld über einen Typ-Knoten fest.
- Sichten Sie die jeweiligen Daten über einen Data Audit-Knoten.
- Bilden Sie das Kündigungsverhalten bestehender Kunden mithilfe eines neuronalen Netzes bestmöglich ab. Verwenden Sie dazu die Trainingsdaten. Vergleichen Sie dabei die Modelle der Trainingsmethoden „Standard“, „Boosting“ und „Bagging“. Für welches Modell entscheiden Sie sich?
- Wenden Sie die erzeugten Modelle auch auf die Validierungsdaten an und evaluieren Sie die Modelle anhand Lift-Chart, Gains-Chart und Fehlklassifikationstabellen.
  Für welches Modell entscheiden Sie sich?

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.1.1 (2/2)

Abwanderungsgefährdete Kunden prognostizieren



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

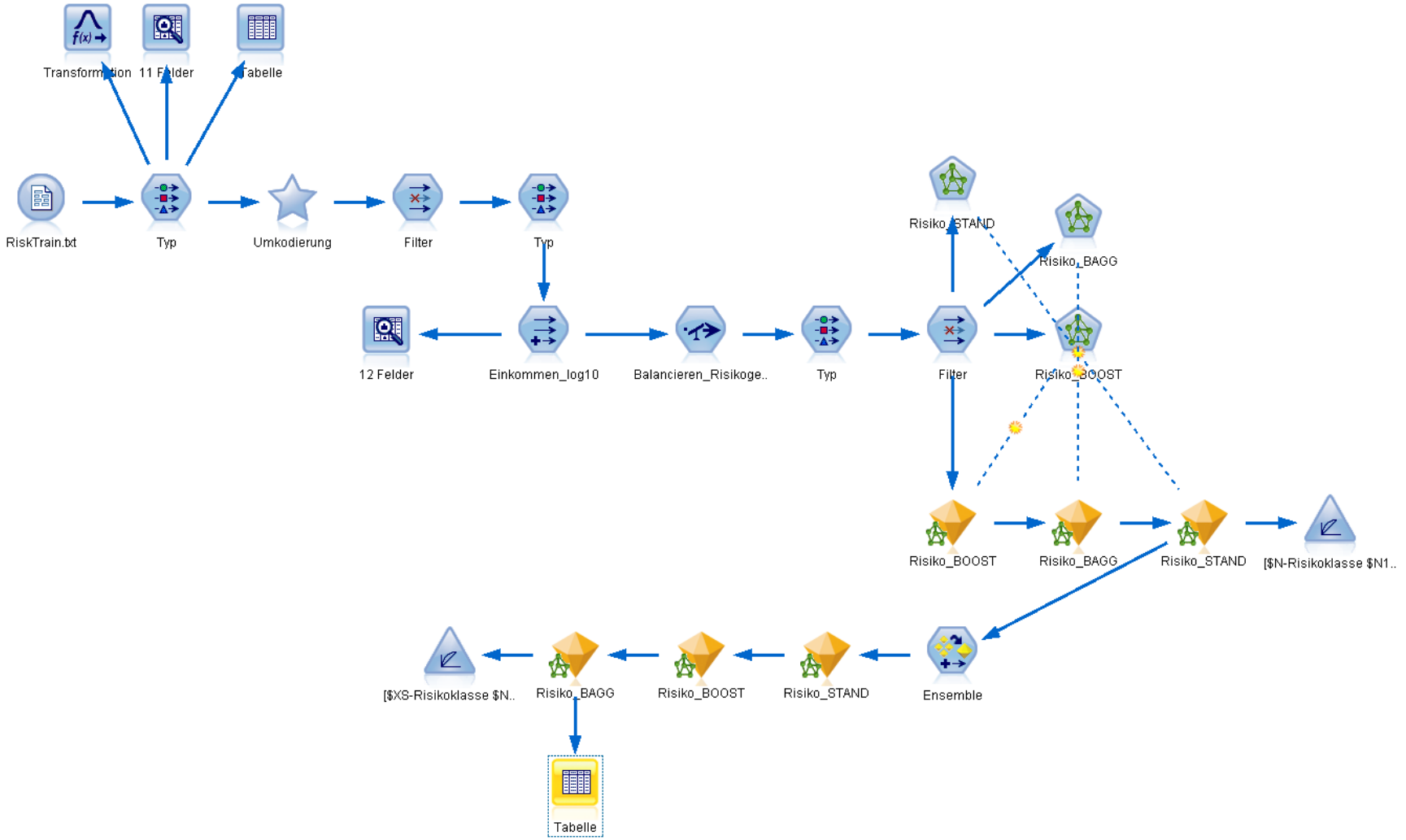
Übung 6.1.2

- **Kundenbonität prognostizieren** (Datei RiskTrain.txt)
  - Ein Kreditinstitut wendet sich an Sie als Experten für neuronale Netzwerke. Sie erhalten streng vertraulich eine Datei mit Kundendaten.
  - Lesen Sie die Daten in Modeler ein und stellen Sie sie mithilfe eines Tabellen-Knotens dar.
  - Im Feld „Risk“ stehen Angaben zum kundenindividuellen Kreditrisiko:
    bad loss: schlechte Bonität und Verlust
    bad profit: schlechte Bonität aber noch Gewinn
    good risk: gute Bonität
  - Entwickeln Sie mithilfe eines neuronalen Netzes ein Modell, dass die Bonität zukünftiger Kunden vorhersagt.
    - Wie könnte ein Stream in Modeler grundsätzlich aussehen?
    - Führen Sie zunächst die Datenaufbereitung durch.
    - Entwickeln Sie in Modeler ein Modell und versuchen Sie es zu optimieren.
    - Wozu könnte ein derartiges Modell in der Praxis eingesetzt werden?
  - Stream speichern unter …. „Übung 6.1.2“

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.1.2

Was fehlt in diesem Stream?



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&R

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

6.4.2 Two-Step

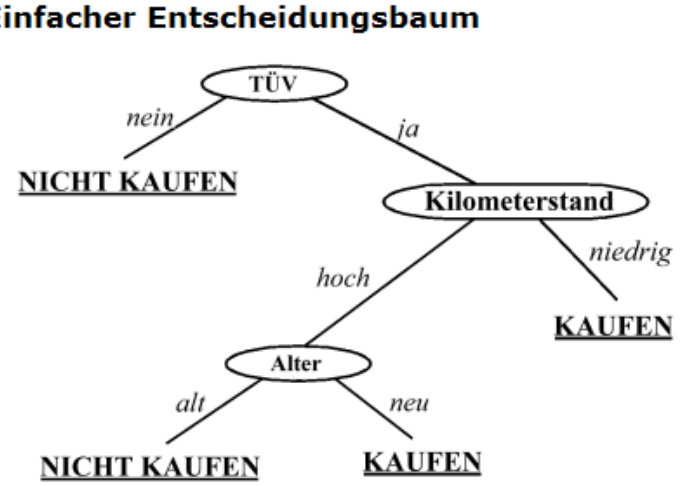
6.5 Assoziationsanalyse

6.5.1 A Priori

Übersicht

- Mithilfe von Entscheidungsbaummodellen werden **Klassifizierungs- bzw. Prognosesysteme** entwickelt, die **zukünftige Beobachtungen auf der Grundlage eines Satzes von Entscheidungsregeln klassifizieren oder vorhersagen**. Wenn beispielsweise Daten in Klassen aufgeteilt sind, die Sie interessieren (z. B. Darlehen mit hohem Risiko im Gegensatz zu Darlehen mit niedrigem Risiko), lassen sich mit diesen Daten Regeln erstellen, die Sie zur Klassifizierung alter oder neuer Fälle verwenden können. So können Sie z. B. einen Baum erstellen, der das Kreditrisiko oder die Kaufabsicht basierend auf Alter und anderen Faktoren klassifiziert.

- Ein Vorteil von Entscheidungsbäumen ist, dass die Argumentationskette hinter dem Modell deutlich wird und sich die Ergebnisse so **gut kommunizieren** lassen. Dies steht im Gegensatz zu anderen "Blackbox"-Modellierungstechniken, bei denen die interne Logik nicht so leicht zu durchschauen ist.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

- Modeler unterstützt vier Methoden zum Erstellen von Entscheidungsbäumen:
  - **C5.0**
  - **C&RT (Classification and Regression Trees)**
  - **CHAID**
  - **Quest**

- Unterschiede der Methoden:

	C5.0	C&RT	CHAID	QUEST
Art des Baumes (Split)	Klassifikationsbaum	Binärbaum	Nichtbinärer Baum	Binärbaum (schneller als C&RT)
Skalenniveau der Zielgröße	Nominal	Beliebig (Nominal und Stetig)	Beliebig (Nominal und Stetig)	Prädiktoren beliebig (Nominal und Stetig), Zielfeld Nominal

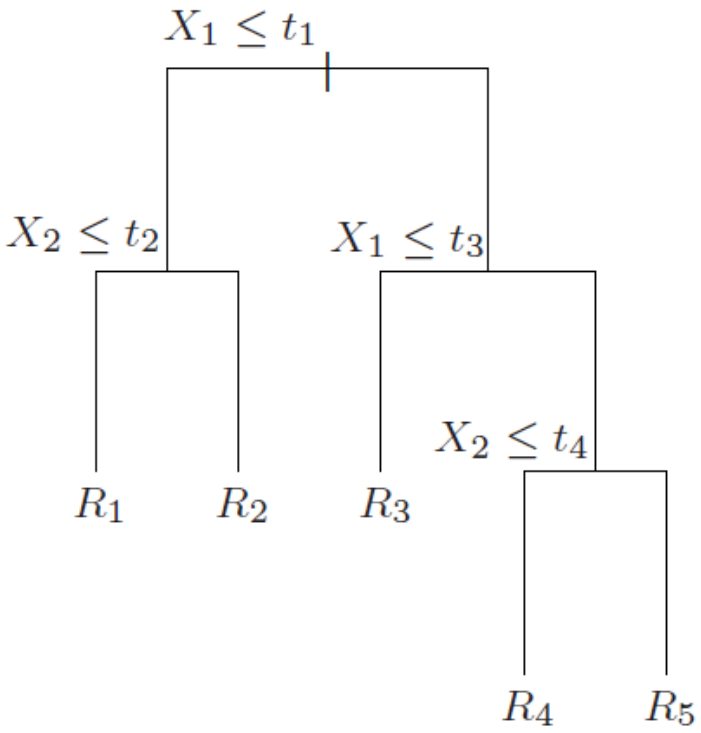
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

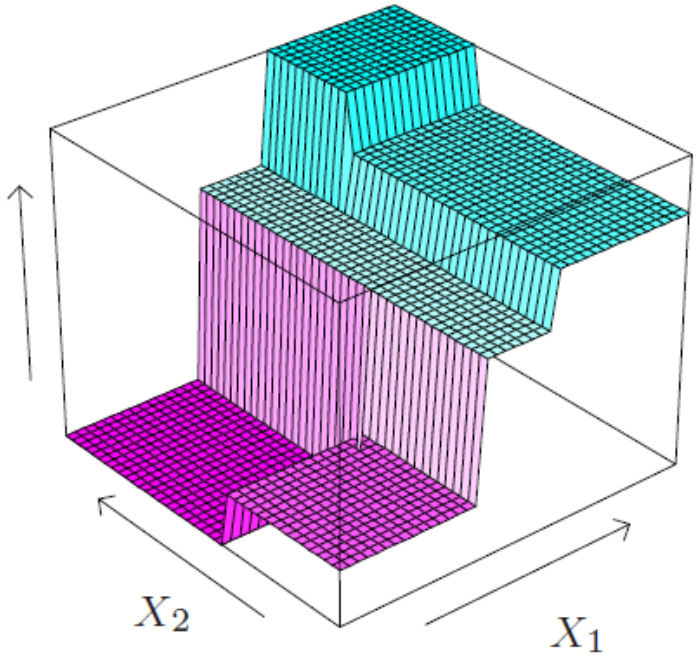
Mögliche Fragestellung

- Welche Eigenschaften meiner bestehenden und früheren Kunden lassen auf erhöhte **Kündigungsbereitschaft** schließen?
  - Ein Entscheidungsbaum versucht jeden Kunden einer Klasse zuzuordnen (Ausprägungen der Zielvariablen).
  - Das Modell untersucht automatisch die gesamte Datenmenge, nach den am besten zur Unterscheidung geeigneten Datenfeldern.
  - Dadurch wird eine Menge an Regeln erzeugt:
    IF (Alter < 36) AND (GESCHLECHT = M) AND ... THEN (KLASSE = KÜNDIGER)
  - Mit anderen Verfahren (z. B. Neuronale Netze) lassen sich vergleichbare Ergebnisse erzielen. → Ausprobieren und vergleichen!
  - Vorteil von Entscheidungsbäumen: Strukturierte und einfach zu interpretierende Darstellung.
- Nachdem das Modell trainiert ist und gelernt hat anhand der Merkmale Kunden zu klassifizieren, kann es genutzt werden, um die Kündigungsbereitschaft neuer Kunden zu prognostizieren. Basierend auf den Ergebnissen lassen sich bspw. Kündigungspräventionsmaßnahmen ableiten.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

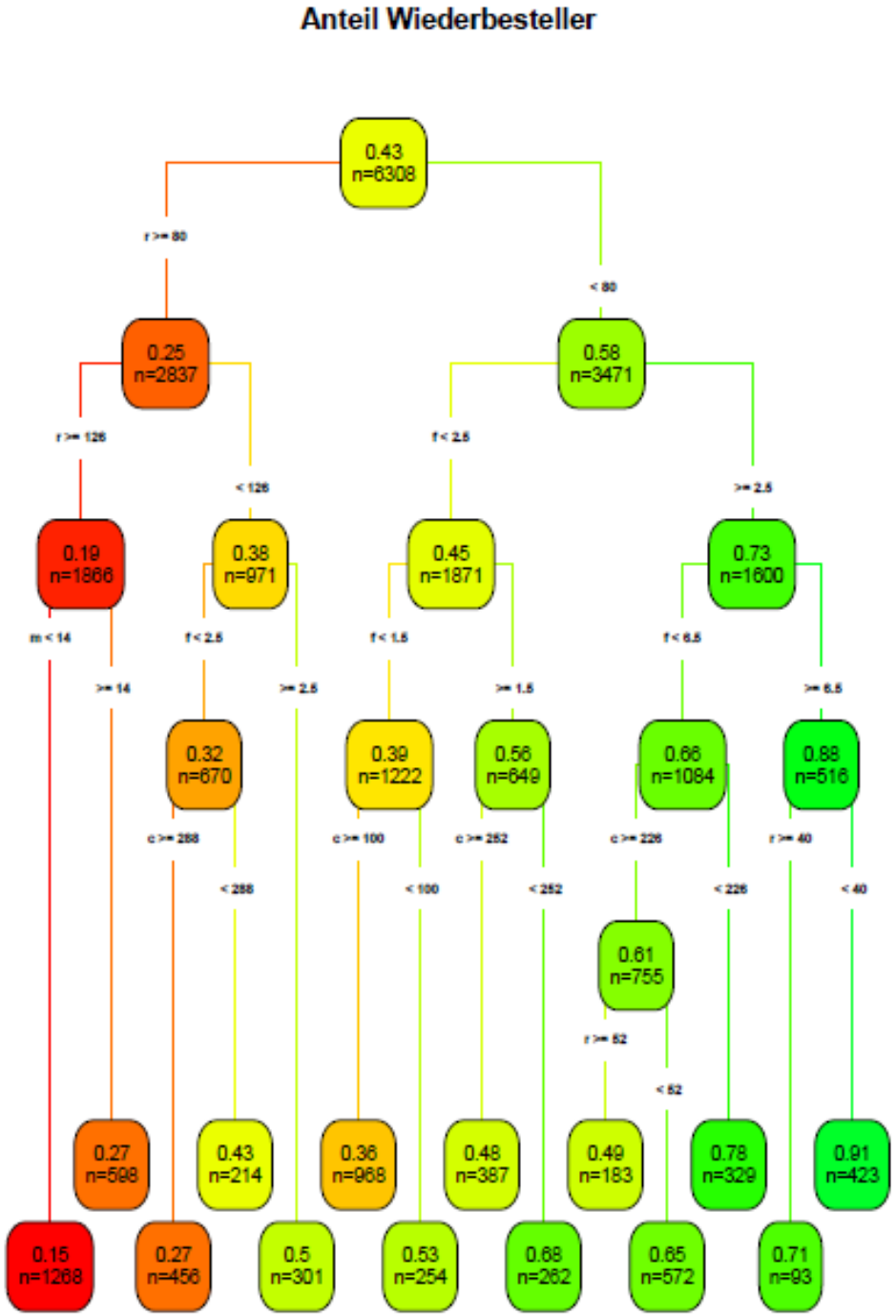
Illustration der Funktionsweise von Regressionsbäumen. Grafik aus „The elements of statistical learning“ von Hastie et al. 2009.





Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde



Anwendung in der Praxis:

Churn-Analyse mit R.

r: Tage seit der letzten Bestellung
f: Anzahl Bestellungen
m: Bisherige Marge des Kunden
c: Kindesalter in Tagen

Alle Werte der unabhängigen Variablen beziehen sich auf einen Zeitpunkt 6 Monate in der Vergangenheit. Geschätzt wird hiermit die Churn-Wahrscheinlichkeit durch Betrachtung des Anteils der Wiederbesteller in den folgenden 6 Monaten (abhängige Variable ist gegeben durch binäre Werte, die kennzeichnen, ob ein Kunde wieder bestellt hat oder nicht).

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&R

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

6.4.2 Two-Step

6.5 Assoziationsanalyse

6.5.1 A Priori

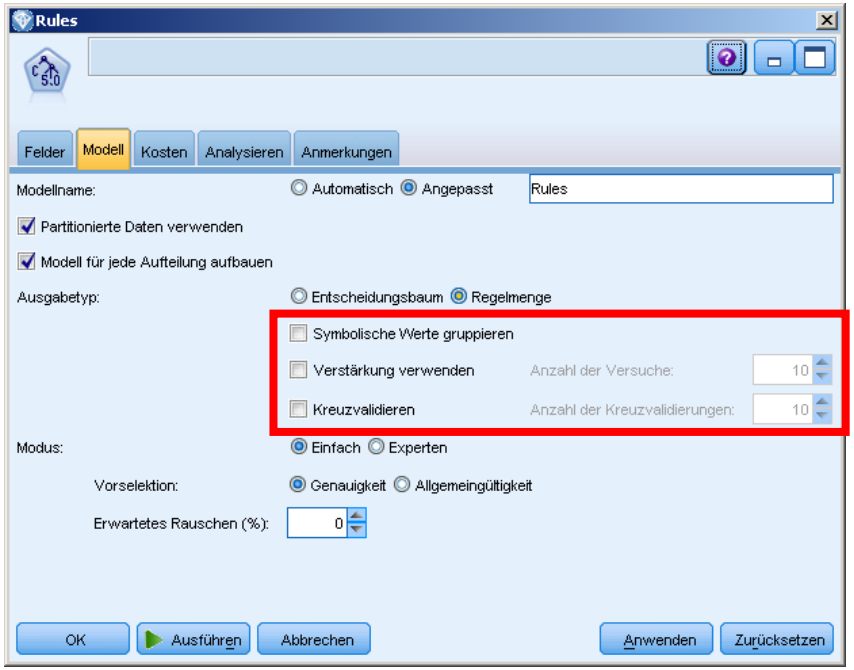
C5.0

- **Voraussetzung**: mindestens ein nominales Zielfeld (Output) sowie mindestens ein Prädiktor-Feld (Input).
- **Vorgehen:** Ein C5.0-Modell teilt die Stichprobe auf der Basis des Feldes auf, das den maximalen Informationsgewinn liefert. Jede durch die erste Aufteilung definierte Unterstichprobe wird dann wieder aufgeteilt - üblicherweise auf der Grundlage eines anderen Feldes. Das Verfahren wird so lange fortgesetzt, bis die Unterstichproben nicht weiter aufgeteilt werden können. Zum Schluss werden die Aufteilungen der untersten Ebene noch einmal untersucht, wobei solche entfernt oder reduziert werden, die nicht wesentlich zum Wert des Modells beitragen.
- **Stärken**: C5.0-Modelle verhalten sich bei Problemen mit fehlenden Daten und einer großen Anzahl von Eingabefelder sehr robust. Sie benötigen für die Schätzung in der Regel keine langen Trainingsphasen.
- **Varianten:** Entscheidungsbaum vs. Regelmenge

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

C5.0

- **Grundeinstellungen:**

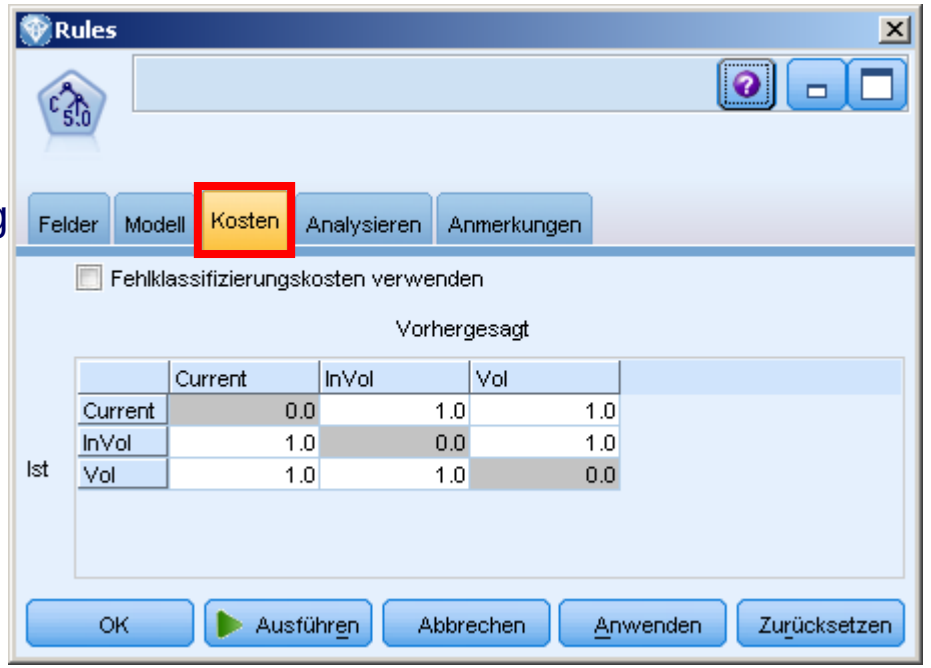


  - **Symbolische Werte gruppieren**
    Bei Aktivierung versucht der C5.0-Alg. Werte mit ähnlichen Mustern in Bezug auf das Output-Feld zu kombinieren.
    Beispiel: Im Feld „Farbe“ existieren die Werte blau, grün und rot. Normalerweise wird für jede Farbe ein Ast gebildet. Sind die Eigenschaften der roten und grünen Objekte sehr ähnlich, werden diese zu einem Ast zusammengefasst.
  - **Verstärkungen verwenden**
    Eine Eigenschaft von C5.0 zur Verbesserung der Genauigkeit. Dabei werden mehrere Entscheidungsbäume erzeugt, wobei jeweils Fehlklassifizierungen des Vorgänger-Baums im Fokus stehen. Anzahl der Modelle lässt sich einstellen - hohe Anzahl verbessert Genauigkeit, erhöht jedoch die Rechenzeit.
  - **Kreuzvalidieren**
    Wenn zu wenig Datensätze zur Bildung von getrennten Trainings- und Validierungsmengen vorliegen, wird das Gesamt-Modell mit kleinen Teilmengen kreuzweise validiert.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

C5.0

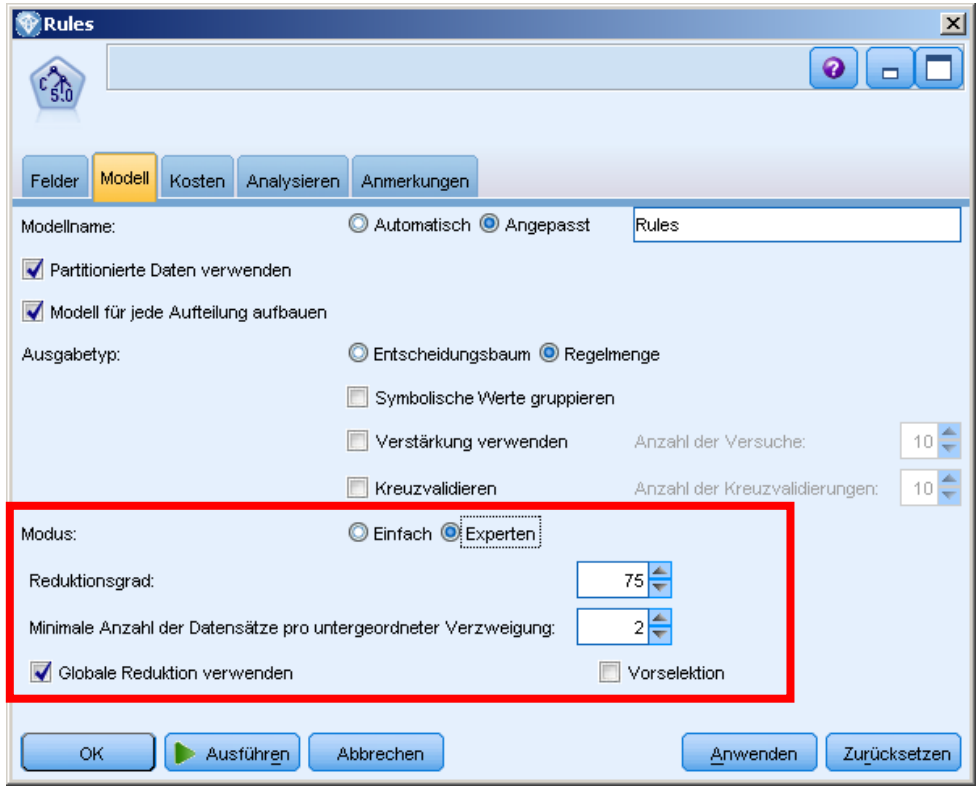
- **Fehlklassifizierungskosten:**
  - Verschiedene Fehlklassifizierungen sind nicht immer mit den gleichen Kosten verbunden – bspw. ist es meist teurer ein hohes Kreditrisiko als niedrig einzustufen als umgekehrt.
  - Alle Kosten sind voreingestellt bei 1.0 (gleiche Kosten) und können in der Matrix geändert werden.
  - Manuelle Änderungen sind nicht automatisch symmetrisch: Wenn die Kosten einer Fehlklassifikation von A als B auf 2.0 gesetzt werden, bleiben die Kosten einer Fehlklassifizierung von B als A weiter bei 1.0.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

C5.0

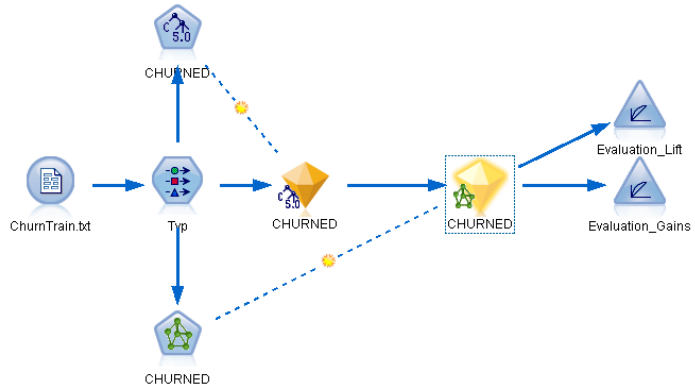
- **Experteneinstellungen:**
  - **Reduktionsgrad**
    Bestimmt die Stärke des Prunings. Ein höherer Wert erzeugt einen kleineren Baum, ein niedriger einen größeren und detaillierteren Baum. Werte zwischen 0 und 100 sind einstellbar - voreingestellt ist 75.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

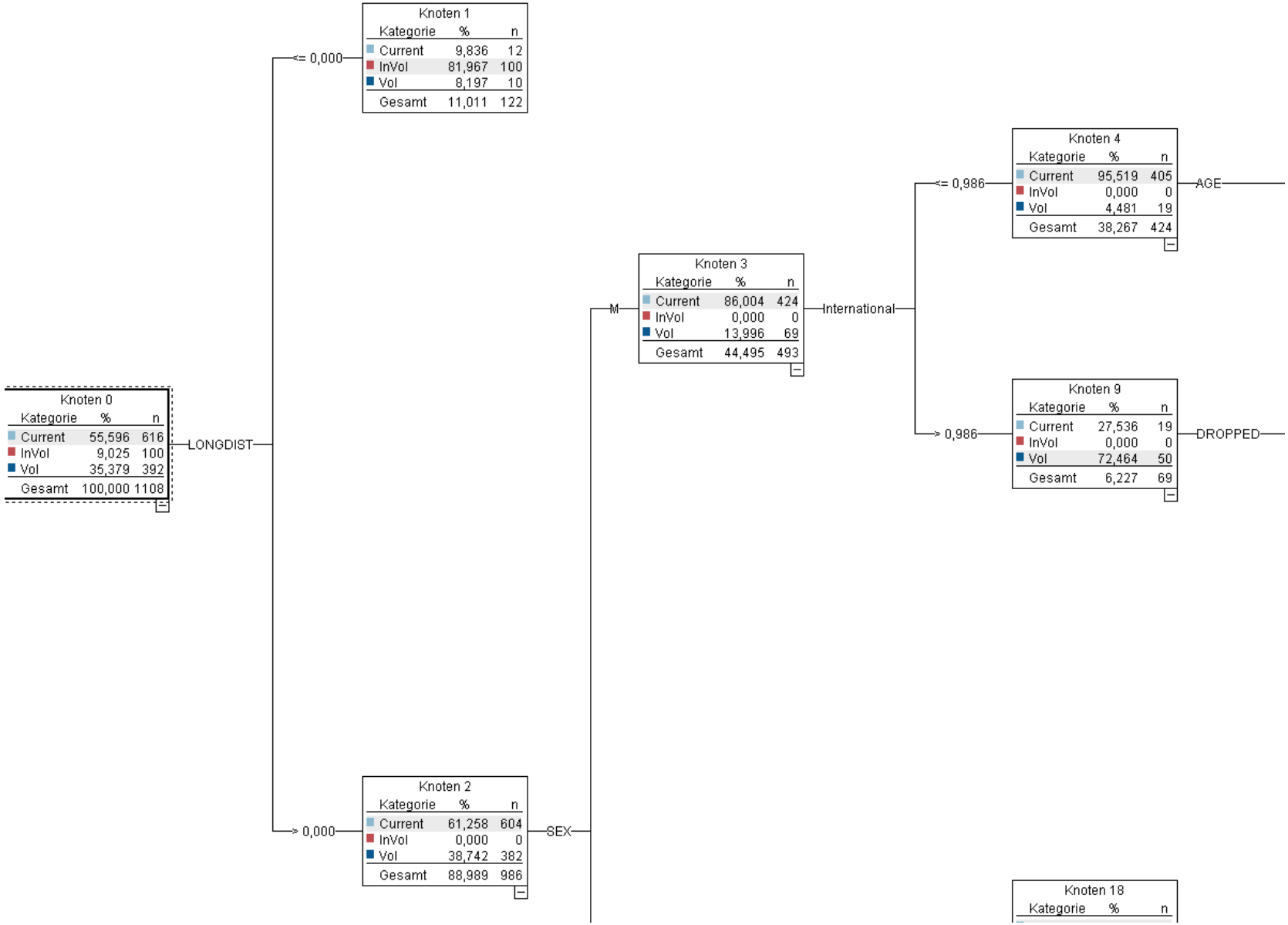
Übung 6.2.1

- **C5.0-Entscheidungsbaum erstellen** (Datei ChurnTrain.txt)
  - Entwickeln Sie ein C5.0-Modell mit der Input-Datei ChurnTrain.txt und der Zielvariable CHURNED. Versuchen Sie das Ergebnis zu interpretieren.
  - Vergleichen Sie das erhaltene Ergebnis mit dem Resultat des zuvor in Übung 6.1.1 erstellten neuronalen Netzes. Verwenden Sie dazu den Auswertungs-Knoten.
  - Welche Empfehlungen können in Bezug auf unsere Ausgangsproblemstellung (Kündigungsprävention) an die Marketing-Abteilung gegeben werden?
  - Was versteht man unter einer „Regelmenge“ und wie kann man diese erstellen?
  - Stream speichern unter … „Übung 6_2_1“



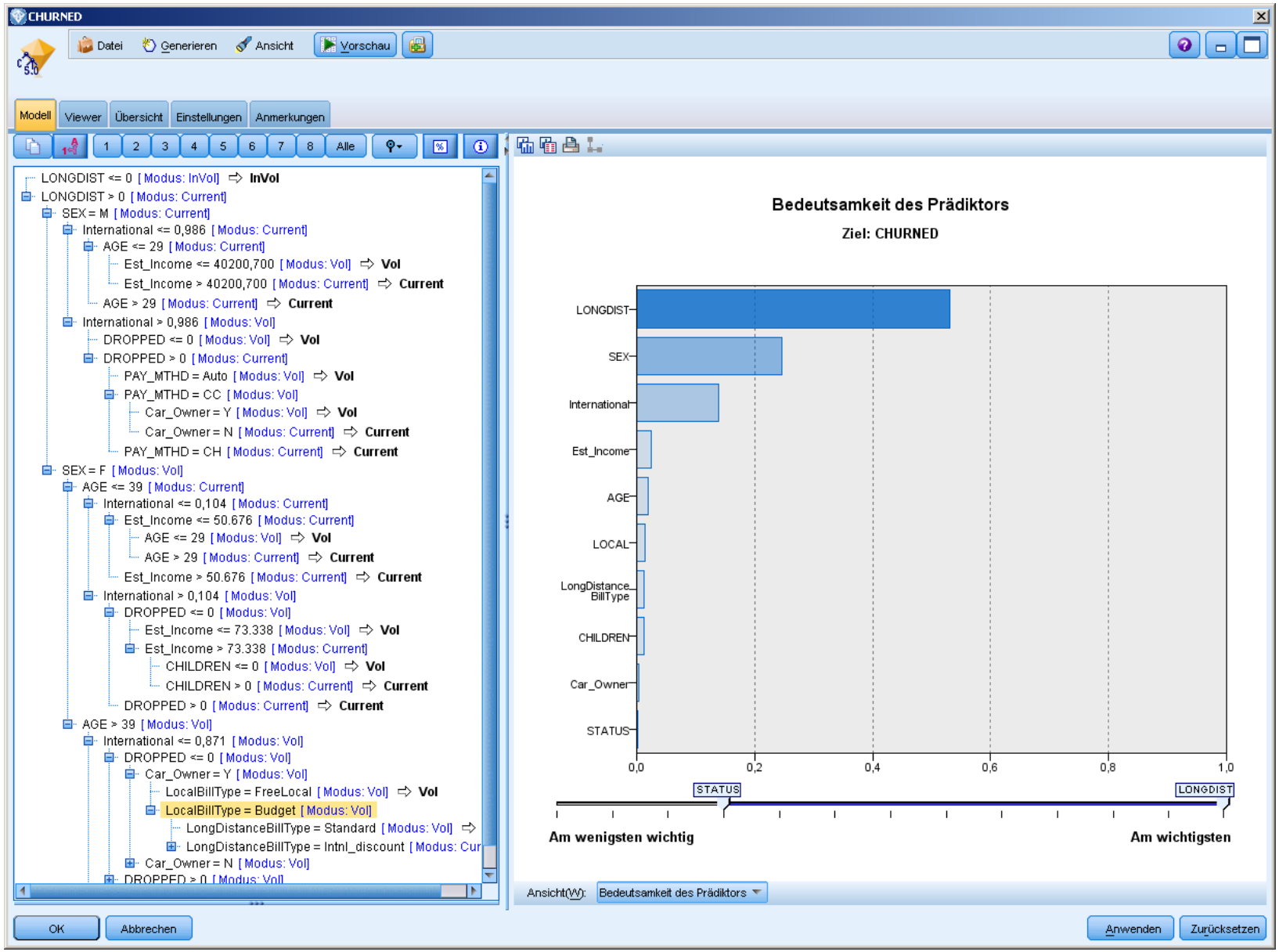
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

C5.0



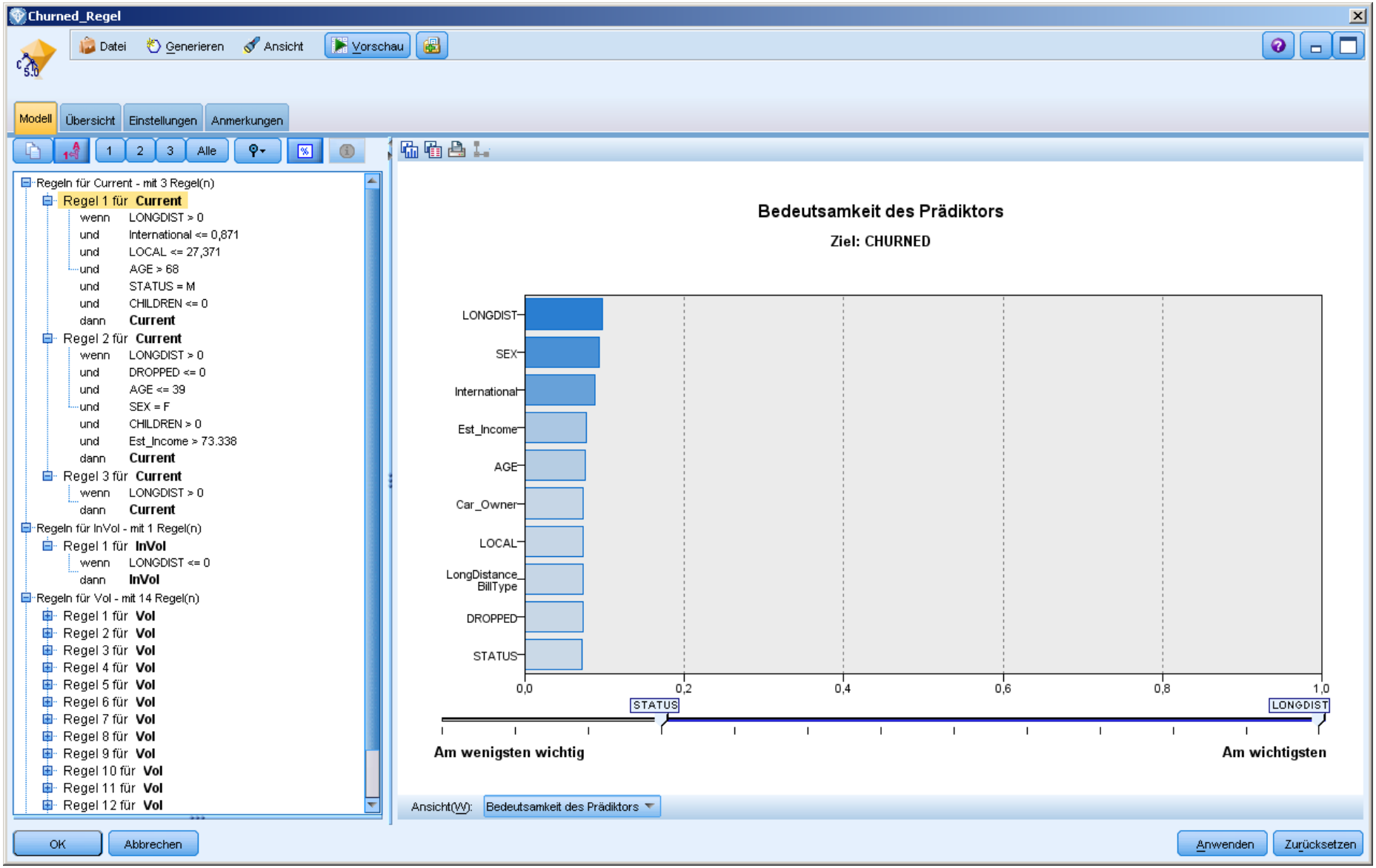
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

C5.0



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

C5.0



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&R

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

6.4.2 Two-Step

6.5 Assoziationsanalyse

6.5.1 A Priori

C&RT

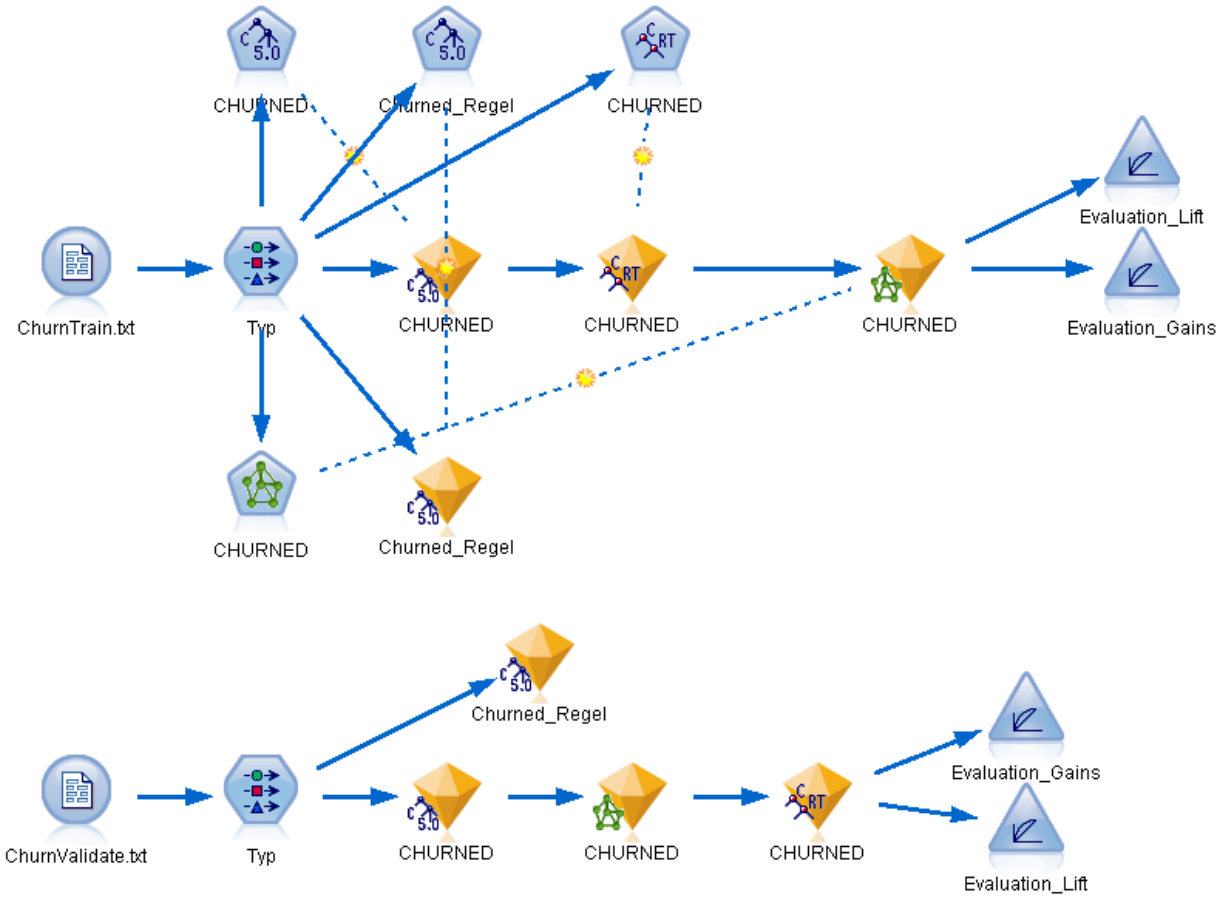
Der **Klassifizierungs- und Regressionsbaum-Knoten** ermöglicht eine Entscheidungsbaum gestützte Klassifikation und Prognose. Ähnlich wie das C5.0 Verfahren benutzt C&R eine rekursive Partitionierung um den Trainings-datensatz in Segmente mit ähnlichen Werten in den Output-Feldern aufzuteilen.

- Bei den Grundeinstellungen des C&R-Trees ist nur die maximale Baumtiefe einstellbar. Die Experten-Einstellungen hingegen sind sehr differenziert und werden im Rahmen der Übung nicht behandelt.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.2.2

- **C&RT-Entscheidungsbaum erstellen** (Datei ChurnTrain.txt)
  - Fügen Sie dem bestehenden Stream ein C&R-Entscheidungsbaum-Modell hinzu und vergleichen Sie die Ergebnisse.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&R

6.3 Lineare Regression

6.4 Cluster-Analyse

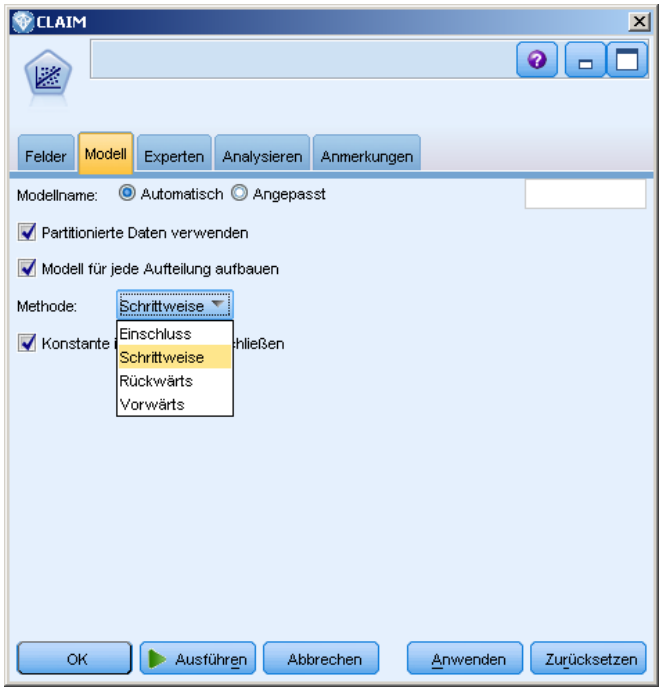
6.4.1 K-Means

6.4.2 Two-Step

6.5 Assoziationsanalyse

6.5.1 A Priori

Einstellungen im Modeler

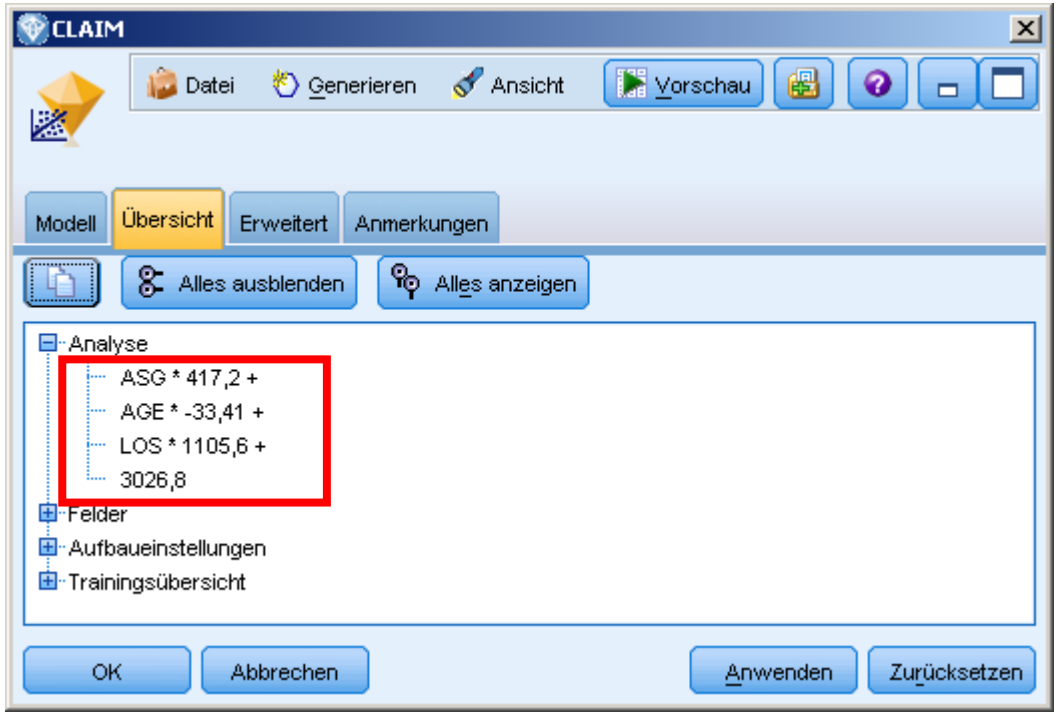


- **Variablen-Auswahl:**
  - **Einschluss (voreingestellt)**
    Alle Input-Datenfelder werden in das Modell aufgenommen. Dies ist vor allem bei einer geringen Anzahl an Merkmalen sinnvoll. Bei einer größeren Anzahl an Variablen wäre eine Selektion durch Experten-wissen möglich.
  - **Schrittweise**
    Schritt für Schritt werden jeweils die Variablen hinzugefügt, die nach statistischen Kriterien die besten Vorhersagewerte liefern. Nach jeder Ergänzung werden die bereits enthaltenen Variablen abermals überprüft und ggf. wieder entfernt. Dieser Prozess wird fortgesetzt, bis keine weitere Verbesserung mehr stattfindet.
  - **Vorwärts**
    Identisches Vorgehen wie bei Schrittweise, allerdings keine Überprüfung und nachträgliche Eliminierung
  - **Rückwärts**
    Zunächst werden alle Input-Variablen übernommen und dann die am wenigsten signifikanten entfernt.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Einstellungen im Modeler

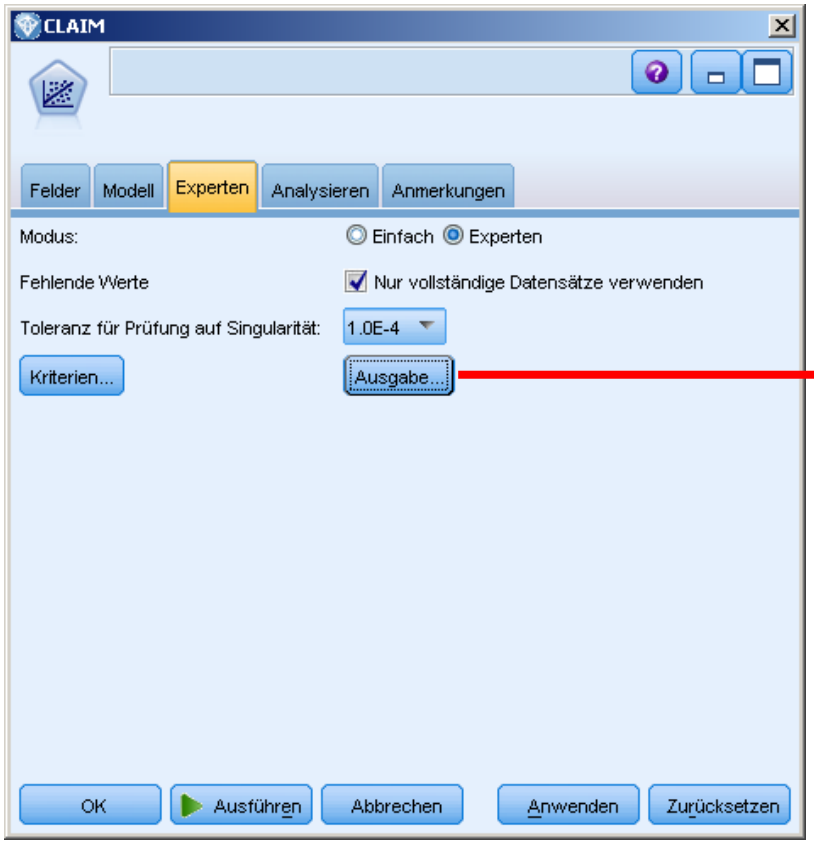
- Der Ergebnis-Nugget enthält die berechnete Regressionsgleichung:

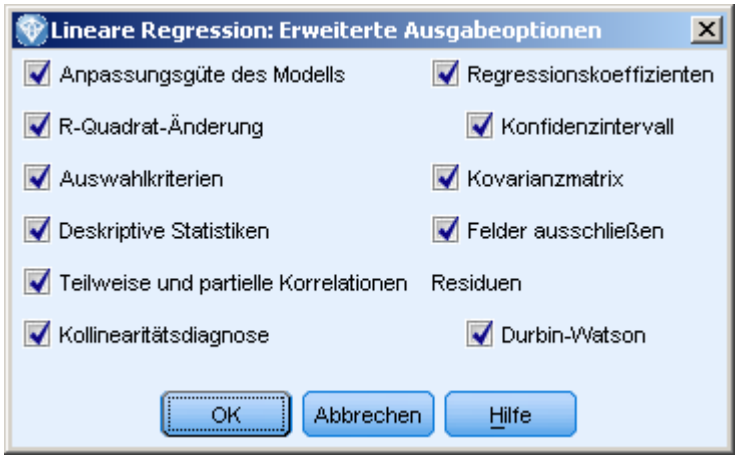


Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Einstellungen im Modeler

- Weitere Aussagen, insbesondere zur Modellgüte lassen sich im **Experten-Modus** einstellen.





Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Interpretation - Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,564(b)	,318	,311	1055,835331

a. Abhängige Variable: CLAIM

b. Einflußvariablen : (Konstante), LOS, ASG, AGE

- **R**
  repräsentiert die Korrelation zwischen dem Output und der optimalen linearen Kombination der Input-Felder. Je näher der Wert an +/-1 ist, desto besser ist das Modell.
- **R-Quadrat**
  kann interpretiert werden als Anteil der Varianz des Outputs, der durch die Input-Felder vorhergesagt wird (hier nur 31,8%). R-Quadrat wird **Bestimmtheitsmaß** genannt. Je näher R-Quadrat an 1 ist, desto besser.
- **Korrigiertes R-Quadrat**
  berücksichtigt die Parameter in der Gleichung (Zahl der Steuergrößen und Fälle; je mehr Variablen desto größer R). Je komplexer das Modell, desto höher der „Abschlag“ auf das Gütemaß.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Interpretation - ANOVA

ANOVA(a)

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	150010746,870	3	50003582,290	44,855	,000(b)
	Residuen	322173803,002	289	1114788,246		
	Gesamt	472184549,871	292			

a. Abhängige Variable: CLAIM

b. Einflußvariablen : (Konstante), LOS, ASG, AGE

- **ANOVA (= Analysis of Variances)**
  Untersuchung, ob das betrachtete Regressionsmodell insgesamt signifikant ist. Die Mittel der Quadrate der Regression („erklärte Varianz“) sind im abgebildeten Beispiel deutlich höher als die Mittel der Quadrate der Residuen („nicht-erklärte Varianz“). Mit anderen Worten kann unser Modell mehr Streuung der abhängigen Variable erklären, als es im Unklaren lässt. Der Quotient aus den Quadratsummen „Regression“ und „Gesamt“ ergibt das bereits erwähnte R-Quadrat.
- Je größer der F-Wert bzw. je kleiner der Wert in der Spalte Signifikanz (entspricht der Wahrscheinlichkeit für Insignifikanz) desto „besser“ ist das Modell.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Interpretation - Koeffizienten

Koeffizienten(a)

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%-Konfidenzintervall für B		Korrelationen			Kollinearitätsstatistik	
		B	Standardfehler	Beta			Untergrenze	Obergrenze	Nullter Ordnung	Partiell	Teil	Toleranz	VIF
1	(Konstante)	3026,754	408,590		7,408	,000	2222,564	3830,943					
	ASG	417,194	114,657	,177	3,639	,000	191,525	642,863	,185	,209	,177	,999	1,001
	AGE	-33,406	11,616	-,140	-2,876	,004	-56,268	-10,543	-,118	-,167	-,140	,998	1,002
	LOS	1105,646	103,600	,519	10,672	,000	901,739	1309,553	,517	,532	,519	,998	1,002

a. Abhängige Variable: CLAIM

- In der Tabelle „Koeffizienten“ können die Parameter der Regressionsgleichung näher betrachtet werden.
- In der Spalte Beta stehen die standardisierten Regressionskoeffizienten, die zum **Vergleich der Wichtigkeit** der einzelnen Variablen in einem multiplen Regressionsmodell herangezogen werden.
- Der T-Test prüft die Nullhypothese $\mathbf{H_0: \beta_j = 0}$. Kann die Nullhypothese für einen Regressionskoeffizienten verworfen werden (hoher empirischer t-Wert), dann gilt der Einfluss von X_j auf Y als signifkant (kleiner Wert in Spalte „Signifikanz“).

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.3.1 (1/3)

- **Betrugserkennung** (Datei insclaim.dat)
  - Wir betrachten ein Regressionsmodell, das die Versicherungsansprüche eines Patienten für einen Krankenhausaufenthalt prognostiziert. Damit sollen insbesondere im Rahmen einer Betrugserkennung (Fraud Detection) größere Abweichungen identifiziert werden.
  - Uns stehen zur Modellbildung die Variablen Alter (**AGE**), Aufenthaltsdauer (**LOS** → length of stay) und Schwere der Krankheit (**ASG**) zur Verfügung, wobei der letzte Wert wiederum aus zahlreichen Indikatoren ermittelt wird.
  - Unser Ziel-Wert ist **CLAIM** - die Höhe des Anspruchs.
  - Für den Fall, dass das Modell brauchbare Vorhersagewerte liefert, interessieren uns insbesondere die Patienten, die von dem prognostizierten Wert stark abweichen. Diese Abweichungen müssen nicht zwangsläufig auf Betrug zurückzuführen sein, die Betrugswahrscheinlichkeit ist jedoch erhöht.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.3.1 (2/3)

- **Aufgabe:**
  - Laden Sie die Quelldatei insclaim.dat und untersuchen Sie diese (Tabelle).
  - Setzen Sie dann das entsprechende Skalenniveau für jede Variable und bestimmen Sie Input- und Output-Variablen. Beachten Sie, dass im Feld ASG Werte einer Ordinalskala enthalten sind (je schwerer die Krankheit, desto höher der Wert).
  - Wie wäre zu verfahren, wenn bspw. „schwer“, „mittel“, „leicht“ als Feldinhalt gegeben wären?
  - Fügen Sie einen **Regression-Knoten** an und editieren Sie ihn. Unser Modell soll CLAIM_Regression heißen.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.3.1. (3/3)

- **Betrugserkennung** (Datei insclaim.dat)
  - Um unserem Ziel der Betrugserkennung näher zu kommen, sind weitere Schritte notwendig. Wir ermitteln mit Hilfe eines **Ableiten**-Knotens („DIFF“) die Differenz zwischen den tatsächlichen Kosten (CLAIM) und den prognostizierten Kosten ($E-CLAIM).
  - Da wir vor allem an den großen Abweichungen interessiert sind, sortieren wir die Werte (**Sortieren**-Knoten) und lassen uns das Ergebnis mit einem **Tabelle**-Knoten anzeigen.
  - Interpretieren Sie das Resultat.

Tabelle (6 Felder, 293 Datensätze) #2

Datei Bearbeiten Generieren

	ASG	AGE	LOS	CLAIM	$E-CLAIM	DIFF
1	1	27	2	11187.400	4753.290	6434.110
2	1	25	3	11715.200	5925.747	5789.453
3	1	26	1	6727.000	3681.050	3045.950
4	1	28	3	8782.200	5825.531	2956.669
5	2	35	1	6489.000	3797.594	2691.406
6	0	38	2	6606.600	3968.636	2637.964
7	0	34	3	7679.000	5207.904	2471.096
8	0	29	3	7798.000	5374.931	2423.069
9	0	37	3	7452.200	5107.687	2344.513
10	0	26	2	6475.000	4369.502	2105.498

Tabelle Anmerkungen

OK

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&RT

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

6.4.2 Two-Step

6.5 Assoziationsanalyse

6.5.1 A Priori

Grundlagen

- Clustermodelle konzentrieren sich auf die **Ermittlung ähnlicher Datensätze** und Beschriftung der Datensätze anhand der Gruppe, in die sie gehören.
- Dies geschieht ohne Vorkenntnisse zu Gruppen und ihren Eigenschaften. Möglicherweise ist nicht einmal bekannt, wie viele Gruppen gesucht werden sollen. Hierin liegt der Unterschied der Clustermodelle zu anderen Techniken des maschinellen Lernens, die in Modeler zur Verfügung stehen: es gibt keine zuvor definierte Ausgabe und **kein Zielfeld für das vorherzusagende Modell**. Clustermodelle zählen zum **Unüberwachten Lernen**, da es keinen externen Standard gibt, mit dem die Klassifizierungsleistung des Modells beurteilt werden könnte.
- Im Rahmen der Übung werden zwei Methoden zur Cluster-Analyse behandelt:

K-Means Clustering

Two-Step Clustering

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&RT

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

6.4.2 Two-Step

6.5 Assoziationsanalyse

6.5.1 A Priori

K-Means

K-Means Clustering:
K steht für die vorgegebene Anzahl gewünschter Cluster. „Means” deutet darauf hin, dass der Mittelpunkt eines Cluster jeweils das Cluster repräsentiert. Als Distanzmaß wird die Euklidische Distanz verwendet.

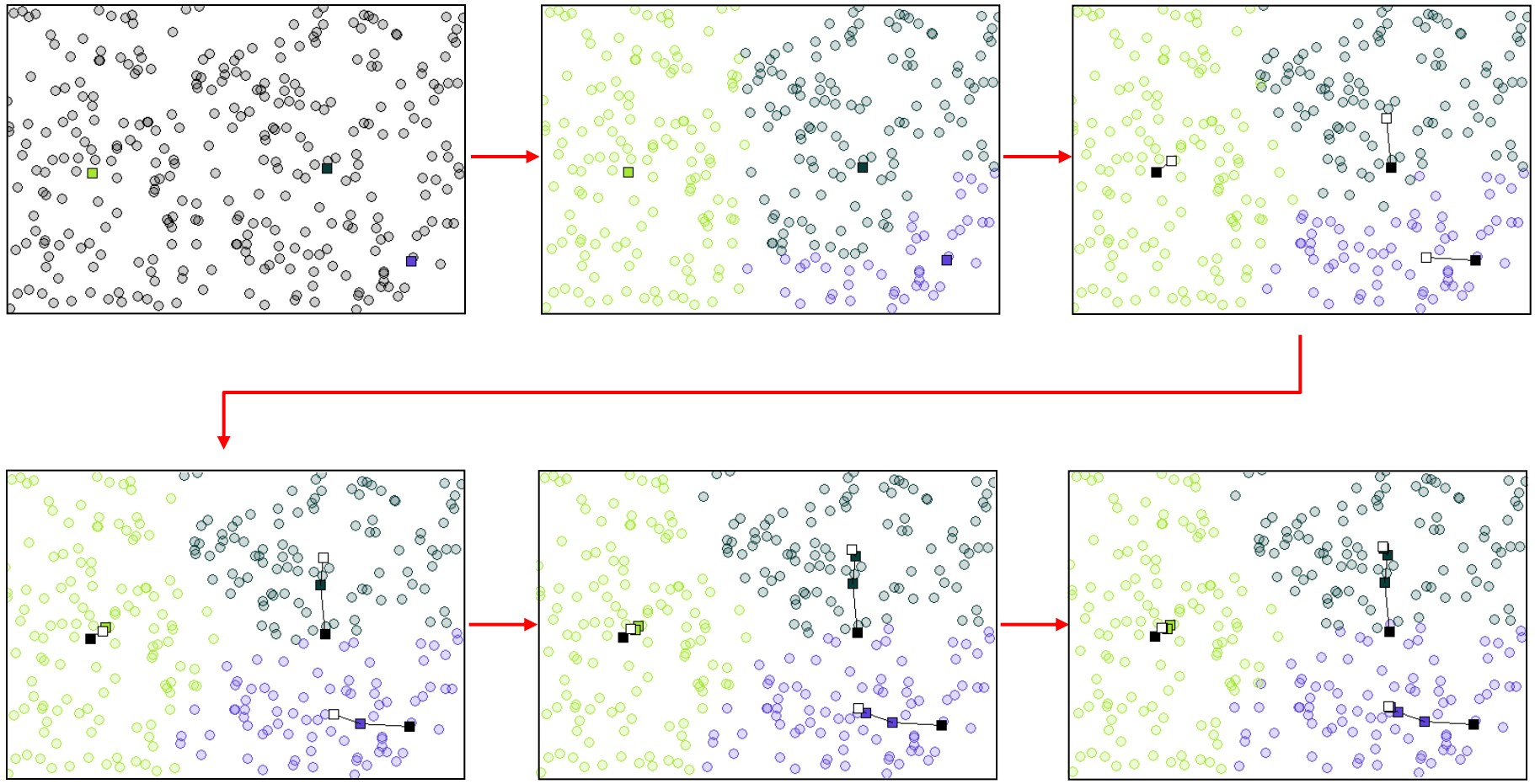
- **Idee:**
  Auf Basis von (mathematischen) Ähnlichkeiten wird iterativ jeder Datenpunkt einem Cluster zugewiesen.

- **Ablauf:**

  1. Festlegen der Clusterzahl
  2. Zufällige Wahl der Clusteranfangszentren
  3. Sukzessive Auswahl aller Datenpunkte und Berechnung der Distanzen zu allen Clusterzentren
  4. Zuweisen zum nächstgelegenen Zentrum
  5. Nach Durchlauf Berechnung der aktualisierten Clusterzentren
  6. Wiederholen bis Abbruchkriterium erreicht ist

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

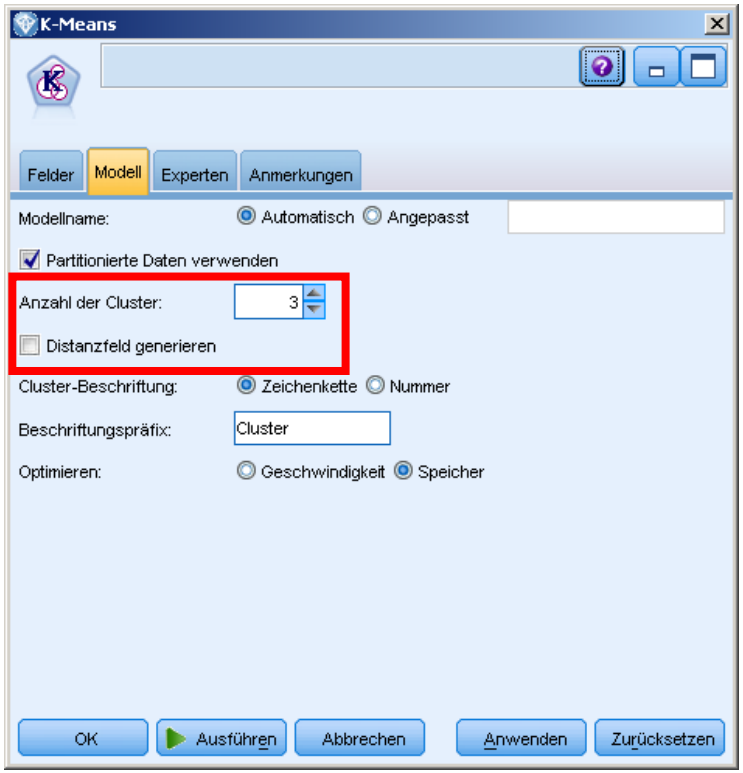
Beispiel



Quelle: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

K-Means



- **Grundeinstellungen:**
  - **Anzahl der Cluster**
    Die Anzahl der Cluster wird vorab eingestellt. Die Default-Einstellung für die Anzahl der Cluster beträgt 5. Je nach Datenlage (z. B. sehr heterogene Daten, sehr große Datenmengen etc.) sollte diese verändert werden.
  - **Option „Distanzfeld generieren“**
    Bei Aktivierung wird ein Datenfeld mit dem Namen $KMD-Kmeans erzeugt, in dem die Distanz zwischen dem spezifischen Objekt und dem dazugehörigen Cluster-Zentrum gespeichert wird.
  - Hinweis zur Datenaufbereitung: K-Means standardisiert Daten selbständig.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

K-Means

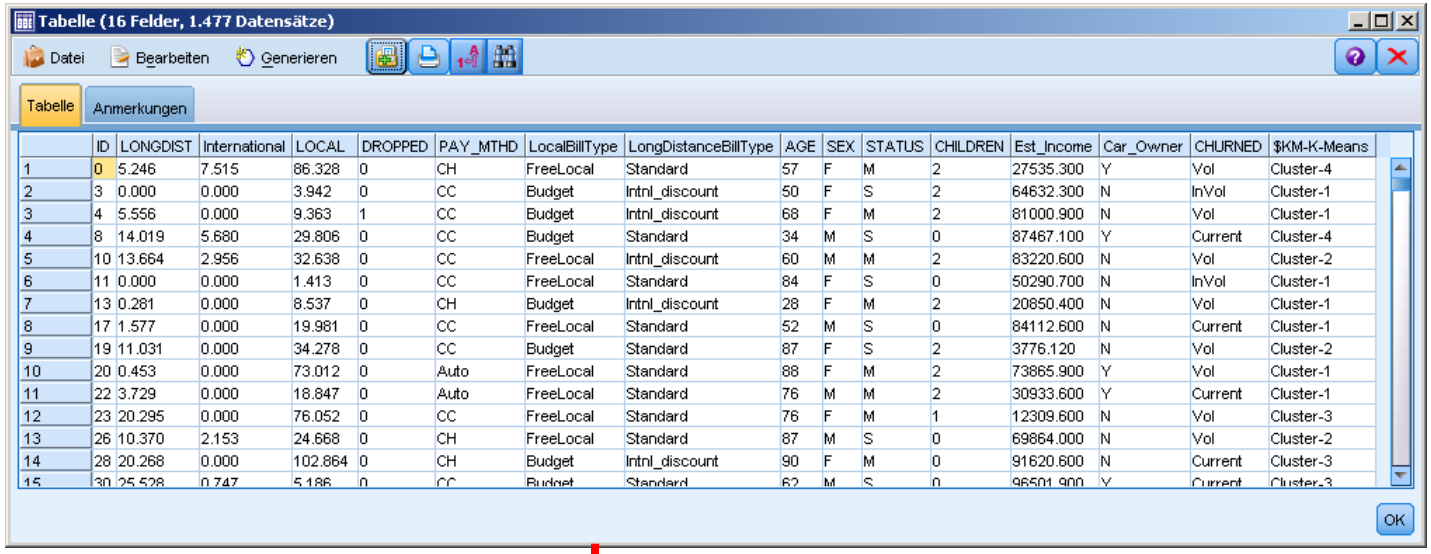
- **Experten-Einstellungen:**
  - **Stopp-Kriterium**
    Der Algorithmus stoppt die iterative Berechnung der Abstände und Distanzmaße wenn ein Stopp-Kriterium erreicht ist. Die Standard-Einstellung führt zu einem Abbruch nach 20 Iterationen oder wenn die größte Änderung der Cluster-Zentren < 0.000001 ist - je nachdem, was zuerst eintritt.

  - **Option „Verschlüsselungswert für Sets“**
    Nominale Datenfelder müssen zur Verarbeitung in mehrere binäre Felder umkodiert werden. Um zu verhindern, dass umkodierte binäre Variablen die Cluster-Lösung dominieren, ist eine Gewichtung von 0.707107 (Wurzel aus 0.5) voreingestellt. Es können Werte zwischen 0.001 und 1.0 ausgewählt werden, wobei Werte unter 0.707 die Bedeutung von binären Variablen verringern, Werte darüber entsprechend erhöhen.

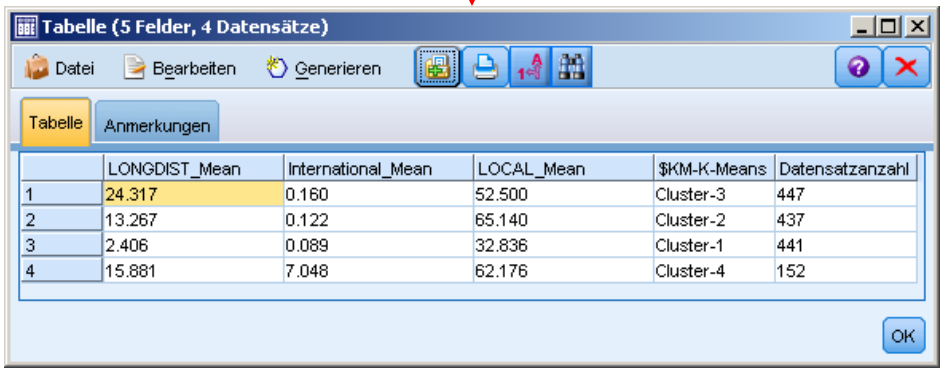
Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

K-Means

- Zur übersichtlichen Darstellung der einzelnen Cluster kann wieder auf den **Aggregieren-Knoten** zurückgegriffen werden.



	ID	LONGDIST	International	LOCAL	DROPPED	PAY_MTHD	LocalBillType	LongDistanceBillType	AGE	SEX	STATUS	CHILDREN	Est_Income	Car_Owner	CHURNED	$KM-K-Means
1	0	5.246	7.515	86.328	0	CH	FreeLocal	Standard	57	F	M	2	27535.300	Y	Vol	Cluster-4
2	3	0.000	0.000	3.942	0	CC	Budget	Intnl_discount	50	F	S	2	64632.300	N	InVol	Cluster-1
3	4	5.556	0.000	9.363	1	CC	Budget	Intnl_discount	68	F	M	2	81000.900	N	Vol	Cluster-1
4	8	14.019	5.680	29.806	0	CC	Budget	Standard	34	M	S	0	87467.100	Y	Current	Cluster-4
5	10	13.664	2.956	32.638	0	CC	FreeLocal	Intnl_discount	60	M	M	2	83220.600	N	Vol	Cluster-2
6	11	0.000	0.000	1.413	0	CC	FreeLocal	Standard	84	F	S	0	50290.700	N	InVol	Cluster-1
7	13	0.281	0.000	8.537	0	CH	Budget	Intnl_discount	28	F	M	2	20850.400	N	Vol	Cluster-1
8	17	1.577	0.000	19.981	0	CC	FreeLocal	Standard	52	M	S	0	84112.600	N	Current	Cluster-1
9	19	11.031	0.000	34.278	0	CC	Budget	Standard	87	F	S	2	3776.120	N	Vol	Cluster-2
10	20	0.453	0.000	73.012	0	Auto	FreeLocal	Standard	88	F	M	2	73865.900	Y	Vol	Cluster-1
11	22	3.729	0.000	18.847	0	Auto	FreeLocal	Standard	76	M	M	2	30933.600	Y	Current	Cluster-1
12	23	20.295	0.000	76.052	0	CC	FreeLocal	Standard	76	F	M	1	12309.600	N	Vol	Cluster-3
13	26	10.370	2.153	24.668	0	CH	FreeLocal	Standard	87	M	S	0	69864.000	N	Vol	Cluster-2
14	28	20.268	0.000	102.864	0	CH	Budget	Intnl_discount	90	F	M	0	91620.600	N	Current	Cluster-3
15	30	25.528	0.747	5.186	0	CC	Budget	Standard	62	M	S	0	96501.900	Y	Current	Cluster-3

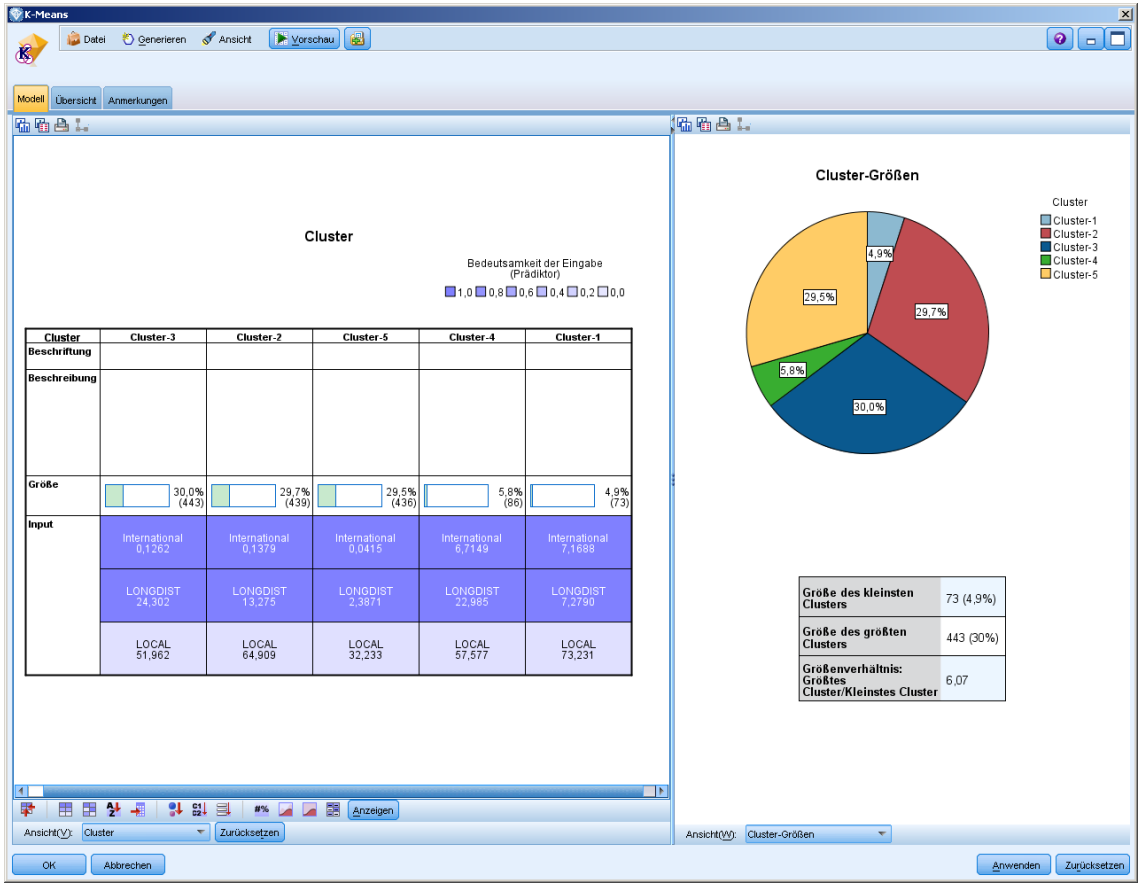


	LONGDIST_Mean	International_Mean	LOCAL_Mean	$KM-K-Means	Datensatzanzahl
1	24.317	0.160	52.500	Cluster-3	447
2	13.267	0.122	65.140	Cluster-2	437
3	2.406	0.089	32.836	Cluster-1	441
4	15.881	7.048	62.176	Cluster-4	152

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.4.1 (1/2)

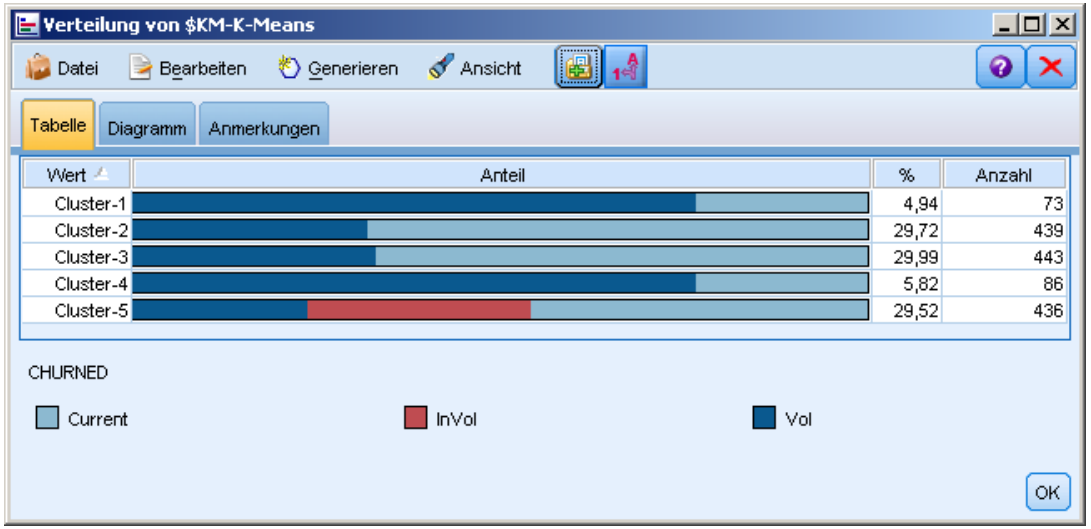
- **K-Means Clustering** (Datei churn.txt)
  - Entwickeln Sie ein Cluster-Modell mit der Zielsetzung, **fünf Segmente für unterschiedliche Kundenbindungsmaßnahmen** zu finden. Verwenden Sie dazu wieder die Datei *churn.txt* mit den drei Input-Variablen Local, Longdist und International.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.4.1 (2/2)

- Neben der inhaltlichen Beschreibung der Cluster ist es im Rahmen unserer Aufgabenstellung interessant, den Zusammenhang zwischen der Variable CHURNED und der Clusterzugehörigkeit zu betrachten.
- **Aufgabe:**
  - Verwenden Sie einen **Verteilung-Knoten** und hängen Sie ihn an Ihr Modell an.
  - Wählen Sie in der List-Box „Feld“ die Variable $KM-Kmeans und als Überlagerung die Variable CHURNED. Aktivieren Sie zusätzlich die Checkbox „Nach Farbe normalisieren“.



Freiwillige Kündiger, die wahrscheinlich interessanteste Gruppe, scheinen in Cluster eins und vier zu fallen. Normalerweise sollten an dieser Stelle weitere Modelle mit anderen Variablen getestet werden.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&RT

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

6.4.2 Two-Step

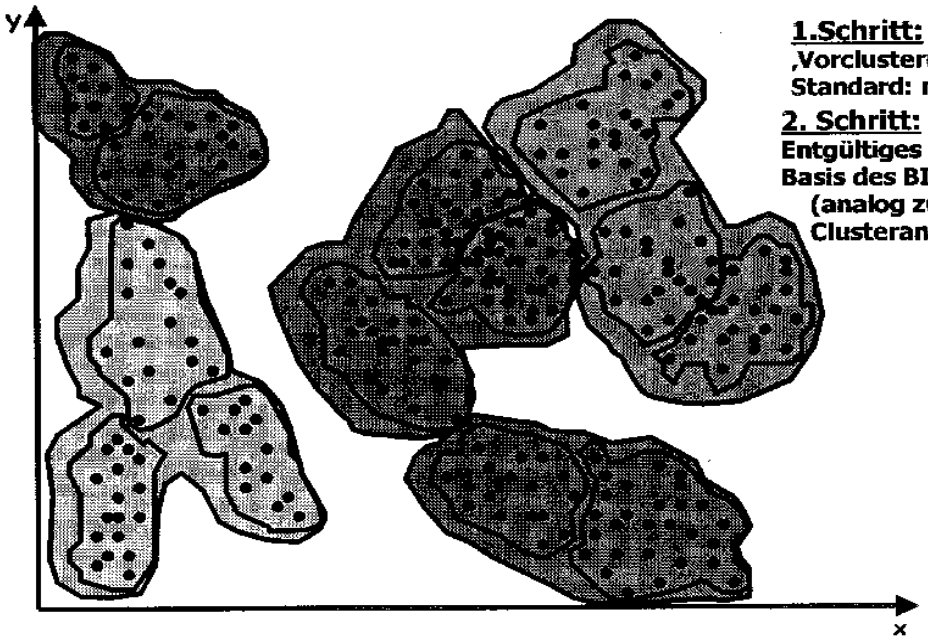
6.5 Assoziationsanalyse

6.5.1 A Priori

Two-Step

Das Two-Step-Clustering eignet sich insbesondere bei großen Datenmengen. Die **optimale Cluster-Anzahl wird automatisch bestimmt**.

- Two-Step besitzt keine Experten-Einstellungen. Daher werden die Werte aller Felder als Default standardisiert (Checkbox). Manuelle Einstellungen zur Gewichtung einzelner Variablen sind nicht möglich.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übung 6.4.2

- **Two-Step Clustering**
  - Öffnen Sie den Stream aus Übung 6.4.1.
  - Führen Sie das Modell durch Anwendung des TwoStep-Knotens.
  - Lassen Sie sich wie bei K-Means die Ergebnisse anzeigen - sowohl im Diamanten, als auch mit Hilfe eines **Aggregieren**-Knotens.

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&RT

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

6.4.2 Two-Step

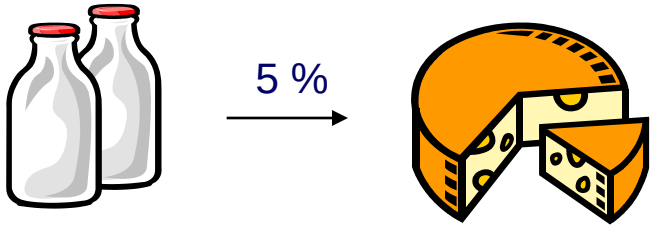
6.5 Assoziationsanalyse

6.5.1 A Priori

Grundlagen

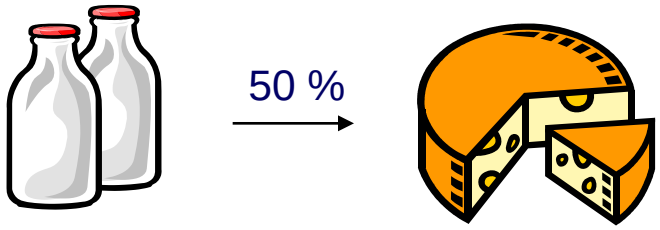
- Die Assoziationsanalyse wird bspw. bei der Untersuchung von Verbundkäufen (Cross Selling) eingesetzt. Die Ergebnisse helfen z. B. bei der Werbeplanung.
- **Definitionen:**

  - **Support (Coverage)**



In jeder 20. Transaktion ist Milch und Käse enthalten

  - **Confidence (Accuracy)**



Wenn Milch gekauft wurde, wurde in 50 Prozent der Fälle auch noch Käse gekauft.

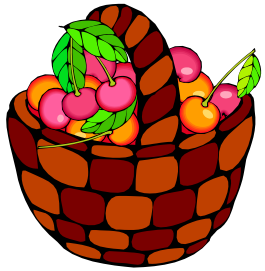
Regelrumpf (condition)

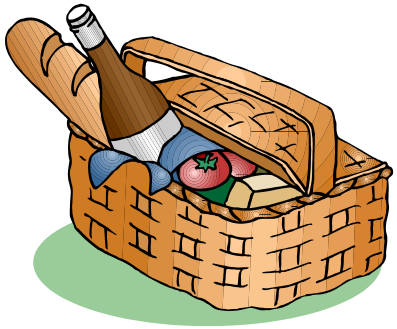
Regelkopf (conclusion)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

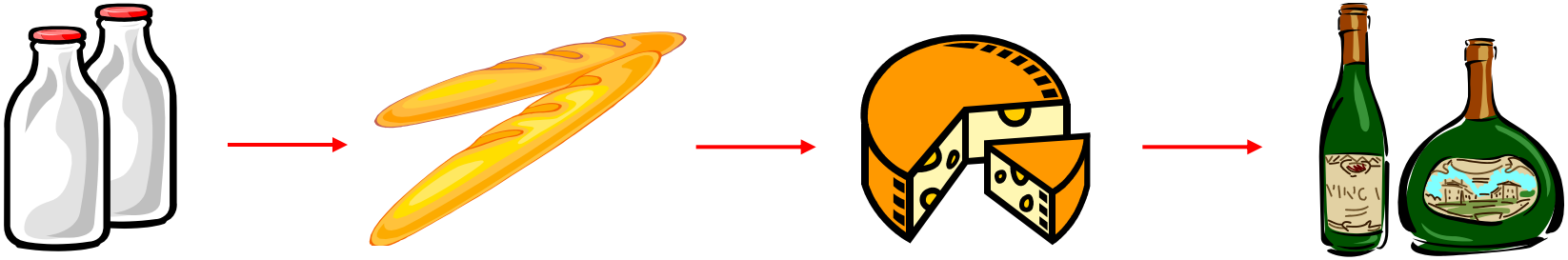
Assoziations- und Sequenzanalyse

Assoziationsanalyse





Sequenzanalyse



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Anwendungsbeispiel

- Modeler-Knoten zur Modellierung von Assoziationsregeln:
  - A Priori
  - CARMA (nicht Bestandteil der Übung)
  - Sequenz (nicht Bestandteil der Übung)

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

Übersicht

6.1 Künstliche Neuronale Netze

6.2 Entscheidungsbäume

6.2.1 C5.0

6.2.2 C&RT

6.3 Lineare Regression

6.4 Cluster-Analyse

6.4.1 K-Means

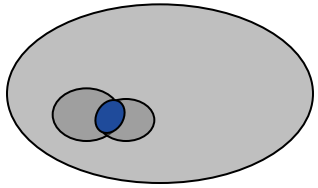
6.4.2 Two-Step

6.5 Assoziationsanalyse

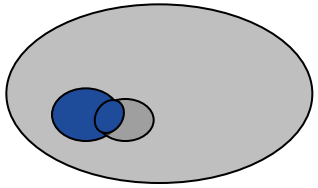
6.5.1 A Priori

A Priori

- **Grundidee:**
  Assoziationsregeln lassen sich durch den Vergleich von Mengen bestimmen.
- **Ablauf:**
  1. Berechnung aller Itemmengen, deren Support über dem Minimum-Support liegen.

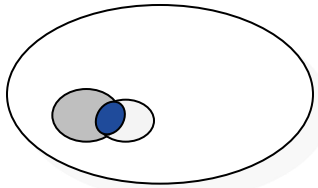


Sequenzanalyse



Assoziationsanalyse

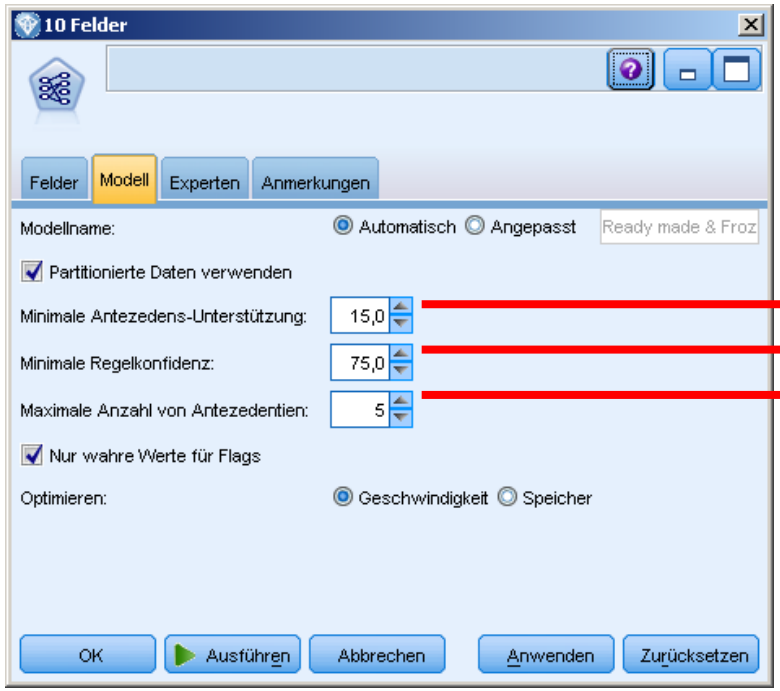
  2. Bildung aller Regeln aus den häufigen Itemmengen, deren Confidence über der Mindest-Confidence liegen.



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

A Priori

- **Voraussetzung:**
  Um eine A Priori-Regelmenge zu erstellen, wird mindestens ein Eingabe- und ein Ziel-Feld benötigt. Ein- und Ausgabefelder (mit der Verwendung Prädiktor, Ziel oder Beides) müssen **binär bzw. nominal skaliert** sein.
- Der A-Priori-Algorithmus arbeitet bei großen Datenmengen im Verhältnis zu den anderen Verfahren schnell. Die Anzahl an Regeln ist unbeschränkt.



Support

Confidence

Maximale Anzahl an Vorbedingungen in den Assoziationsregeln
→ Komplexität der Regeln lässt sich begrenzen

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

A Priori

- Mit dem „A Priori“-Knoten lassen sich Daten in zweierlei Formaten verarbeiten:
  - **Tabellarisch**

Transaktion	Brot	Milch	Zucker
1	J	J	N
2	N	J	N
3	J	J	N
4	N	N	J

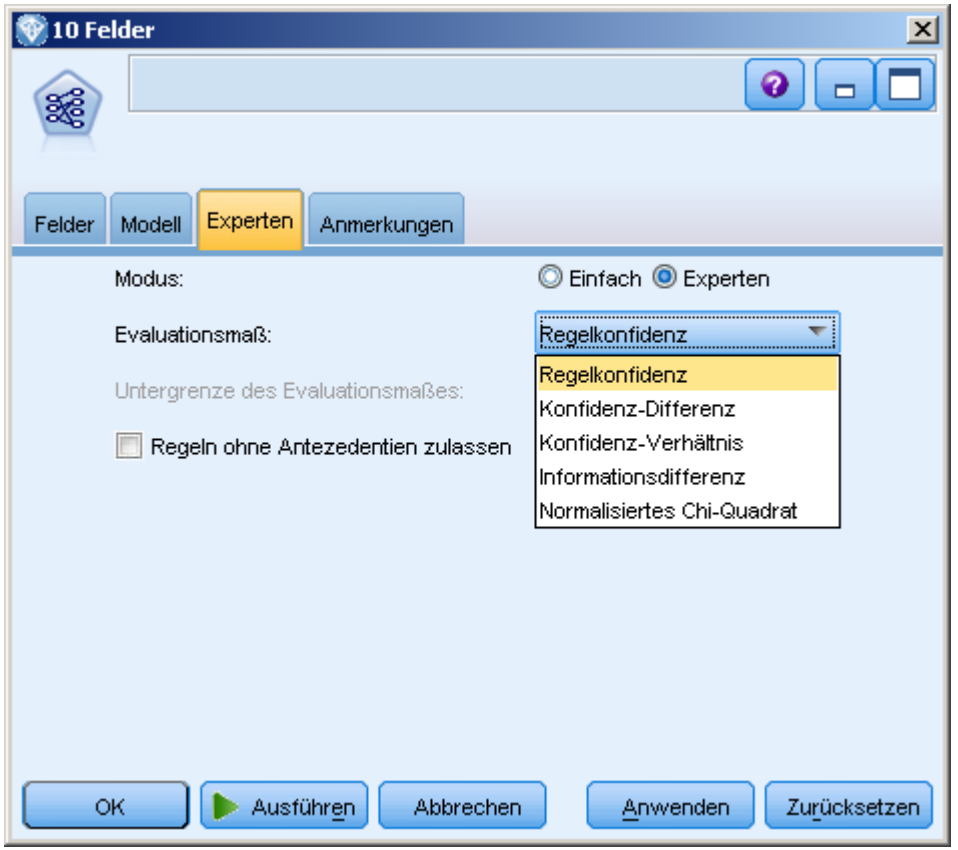
  - **Transaktional:**

Transaktion	Kauf
1	Brot
1	Milch
2	Milch
3	Brot
3	Milch
4	Zucker
...	...

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

A Priori

- **Experten-Einstellungen:**
  Zur Evaluation möglicher Regeln stehen fünf Methoden zur Verfügung:
  - Regelkonfidenz (Standard)
  - Konfidenz-Differenz
  - Konfidenz-Verhältnis
  - Informationsdifferenz
  - Normalisiertes Chi-Quadrat



Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

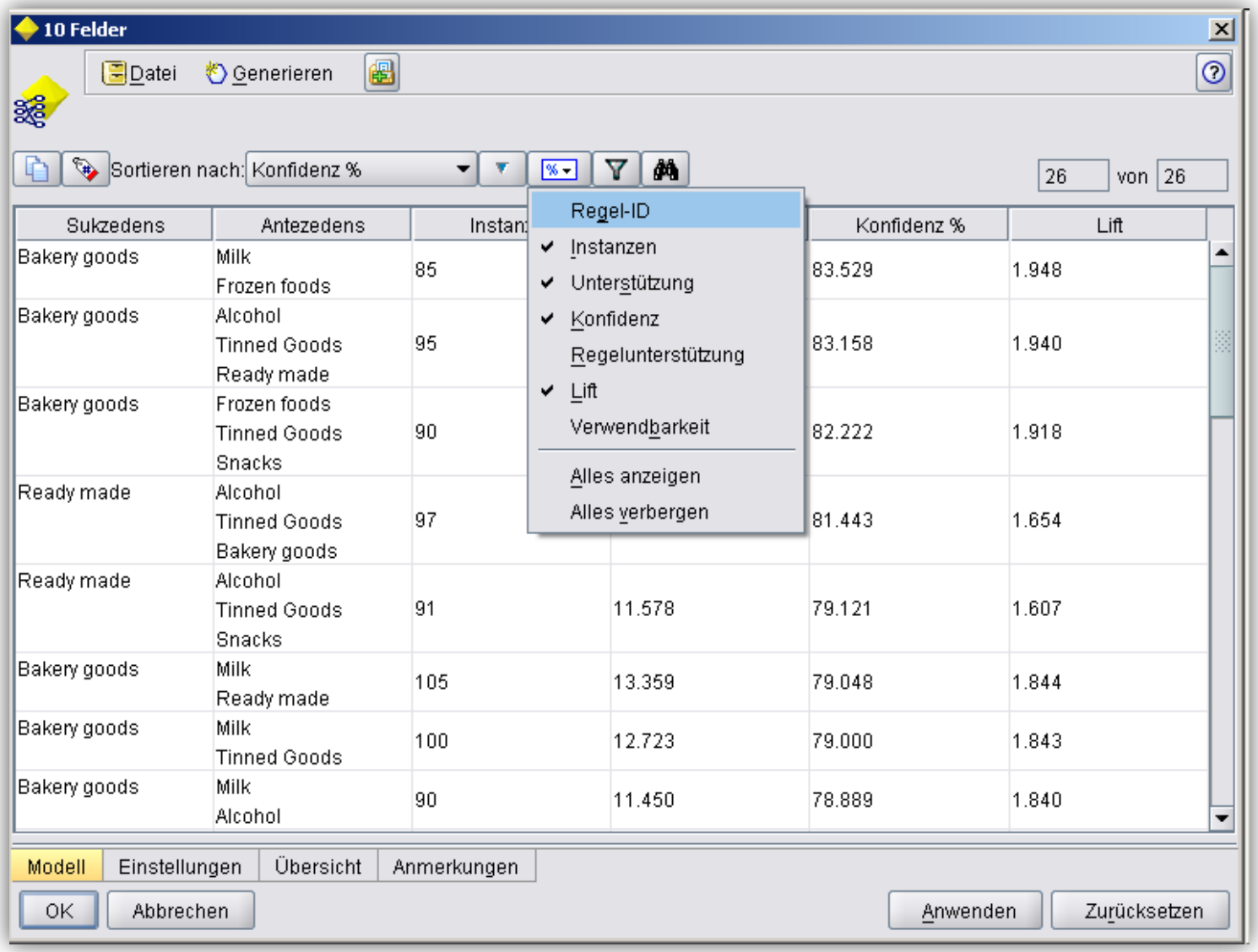
Übung 6.5

A-Priori-Assoziationsanalyse

- Lesen Sie die Datei shopping.txt ein.
- Hängen Sie einen Typknoten an und stellen Sie die Rollen ein. Beachten Sie, dass manche Variablen zugleich Input als auch Output für den „A Priori“-Algorithmus sind.
- Um einen ersten Eindruck zu bekommen, führen Sie den „A Priori“-Knoten ohne Veränderungen aus. Für Mindest-Support und -Konfidenz sind die Werte 10 bzw. 80 eingestellt.
- Führen Sie das Modell aus und betrachten Sie die Daten im **Tabelle**-Knoten.
- Lassen Sie sich die Ergebnisse anzeigen. Wie viele Regeln werden erzeugt? Wie viele Attribute enthalten diese maximal?
- Wie müssen die Einstellungen verändert werden, wenn mehr Regeln angezeigt werden sollen. Erstellen Sie ein neues Modell und interpretieren Sie die Ergebnisse. Was sind „gute“ Regeln und warum?

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde

A Priori



Sukzedens	Antezedens	Instan		Konfidenz %	Lift
Bakery goods	Milk Frozen foods	85		83.529	1.948
Bakery goods	Alcohol Tinned Goods Ready made	95		83.158	1.940
Bakery goods	Frozen foods Tinned Goods Snacks	90		82.222	1.918
Ready made	Alcohol Tinned Goods Bakery goods	97		81.443	1.654
Ready made	Alcohol Tinned Goods Snacks	91	11.578	79.121	1.607
Bakery goods	Milk Ready made	105	13.359	79.048	1.844
Bakery goods	Milk Tinned Goods	100	12.723	79.000	1.843
Bakery goods	Milk Alcohol	90	11.450	78.889	1.840

Katholische Universität Eichstätt-Ingolstadt
Lehrstuhl für ABWL und Wirtschaftsinformatik
Dipl.-Math. Josef Bauer , Vorversion: Dr. Frank Hesse und Prof. Dr. Wilde